

CHANCE AND CHOICE BY  
CARDPACK AND CHESSBOARD





# CHANCE AND CHOICE BY CARDPACK AND CHESSBOARD

AN INTRODUCTION TO PROBABILITY IN  
PRACTICE BY VISUAL AIDS

VOLUME II

by

LANCELOT HOGBEN M.A. (CANTAB.) D.Sc. (LOND.) F.R.S.

PROFESSOR OF MEDICAL STATISTICS IN THE UNIVERSITY OF BIRMINGHAM

LONDON  
MAX PARRISH & CO LIMITED

1955

MAX PARRISH AND CO LIMITED  
55 QUEEN ANNE STREET LONDON WI



PRINTED IN GREAT BRITAIN AT  
THE UNIVERSITY PRESS  
ABERDEEN



# TABLE OF CONTENTS

## CHAPTER 11

PAGE 427

### THE ALGORITHMS OF SUMMARISATION

Introduction—11.01 : Three Types of Grid—11.02 : Algorithms of the Score-Frequency Grid—11.03 : The Independence Grid—11.04 : Tautologies of Correlation—11.05 : Tautologies of the Score-Grid—11.06 : Addition of Covariance—11.07 : Summation by Figurate Series—11.08 : The Generating Function as a Grid Operation.

## CHAPTER 12

PAGE 475

### MODELS OF BIVARIATE UNIVERSES

12.00 : Meaning of the Models—12.01 : The Umpire Bonus Model—12.02 : The Non-Replacement Model—12.03 : The Two-Pack Model—12.04 : The Unit Sample Rectangular Model—12.05 : Lexis Models—12.06 : Sample Size as a Source of Variation—12.07 : The Orthogonal Lottery Model—12.08 : The Standard Score Symbolism—12.09 : Regression and Concomitant Variation.

## CHAPTER 13

PAGE 532

### ASSUMPTIONS UNDERLYING ANALYSIS AND SYNTHESIS OF VARIANCE

13.00 : Analysis of Variance—13.01 : Multiple Criteria of Classification—13.02 : The Complete Sampling Distribution—13.03 : Criteria of Homogeneity—13.04 : A Balance Sheet for Two Criteria—13.05 : The Additive Principle—13.06 : Balance Sheet for Three Criteria—13.07 : One Criterion of Classification—13.08 : Degrees of Freedom.

## CHAPTER 14

PAGE 579

### MORE ABOUT MOMENTS

14.01 : Reality and Rigour—14.02 : Moment Generating Function—14.03 : Factorial Moments—14.04 : The Normal Distribution—14.05 : Moments of the Distribution of the Mean—14.06 : Moments of a Difference Distribution—14.07 : Normal Approximations—14.08 : Sampling from Different Universes—14.09 : The Use of Paired Differences.

## CHAPTER 15

PAGE 634

### SAMPLING DISTRIBUTIONS

15.01 : The Fundamental Distributions—15.02 : Probability Density—15.03 : Functional Transformations of Sample Distributions—15.04 : The Pearson System—15.05 : The Tabulation of the Gamma Function—15.06 : Score Sum and Mean Score of a Sample from a Gamma Universe—15.07 : The Independence Condition—15.08 : The Chi-Square Table.



# TABLE OF CONTENTS

## CHAPTER 16

PAGE 669

### SIGNIFICANCE TESTS FOR ANALYSIS OF VARIANCE

16.01: The Variance Problem—16.02: The Orthogonal Transformation—16.03: Confidence Limits of Variance Estimates—16.04: Degrees of Freedom for Variance Estimates—16.05: The Paired Difference Test for Small Samples—16.06: The Group Mean Difference Test—16.07: Testing the Variance Ratio—16.08: The Correlation Ratio.

## CHAPTER 17

PAGE 712

### SIMPLE REGRESSION AND DISCRIMINATION

17.00: Regression in Real Work—17.01: Principle of the Fixed *A*-Set—17.02: Computation for the Regression Equation—17.03: Unbiased Estimates of Regression Parameters—17.04: Significance of Estimated Regression Parameters—17.05: The Method of Least Squares—17.06: Regression in the Domain of Concurrence—17.07: Partial Regression and Multiple Correlation—17.08: The Discriminant Function.

## CHAPTER 18

PAGE 764

### ELEMENTS OF ANALYSIS OF COVARIANCE AND OF FACTOR ANALYSIS

18.01: Regression as a Standardising Device—18.02: Analysis of Covariance—18.03: Caveat to Analysis of Covariance—18.04: The Concept of Factor Pattern—18.05: Derivation of the Hierarchical Criterion—18.06: Reliability, Attenuation and Communality—18.07: The Single Factor Pattern—18.08: The Bi-Factor Pattern—18.09: Higher Factor Patterns.

## CHAPTER 19

PAGE 804

### SAMPLING IN A FINITE UNIVERSE AND MANIFOLD CLASSIFICATION

19.01: The Hypergeometric Distribution—19.02: Moments of a Score-Sum Distribution—19.03: Chessboard Derivation of the Multinomial Theorem—19.04: Sampling Without Replacement in the Finite Universe of Representative Scoring—19.05: Difference Distribution for Non-Replacement Sampling—19.06: The So-called Chi-Square Statistic—19.07: The Moments of the So-called Chi-Square Statistic.

## CHAPTER 20

PAGE 840

### SECOND THOUGHTS ON SIGNIFICANCE

20.00: Statistical Inference—20.01: Statistical Inspection—20.02: Bayes' Theorem and the Sequential Ratio—20.03: Limitations of the Unique Null Hypothesis—20.04: The Concept of Test Power—20.05: A Sequential Test Procedure—20.06: Estimation and Confidence—20.07: Estimation and the Bayes' Dilemma—20.08: The Show Must Go On.



## CHAPTER 11

# THE ALGORITHMS OF SUMMARISATION

### 11.00 INTRODUCTION

IN VOL. I our main concern has been with problems involving comparison of scores of not more than 2 samples from the same universe or from identical universes. When this is the end in view, the score may be: (a) the mean or total value of a count (e.g. *income*) or measurement (e.g. *height*) attached to each individual of the group; (b) the proportion or number of individuals with a particular attribute. Whenever we have so far used the method of scoring last mentioned, we have done so on the understanding that the constituent individuals of the group are assignable to one of two classes A or not-A. We have not examined the consequences of any null hypothesis when: (a) our definition of sample structure entails a specification of individuals assignable to more than 2 classes; (b) our comparison of samples involves more than two of them at a time. Such are the issues we shall chiefly explore in this volume.

In doing so, we shall acquaint ourselves with different statistical techniques which have much in common, though their uses may be greatly different. What is common ground may be referable to the logic of classifying and summarising the relevant data or to the algebra of the sampling distributions invoked as a basis for significance tests. For that reason, we shall not attempt to deal comprehensively with any one of them as a separate entity in the chapters which follow; and a brief summary of the lay-out of this volume may therefore be helpful to the beginner who wishes to make the best use of it.

The only theoretical sampling distribution we have so far employed as a basis for a significance test is the *normal*; but we have seen that the reasons for relying on the normal distribution in one context may be quite different from the reasons for relying on it in another. For instance, different reasons justify our belief that the normal curve gives in practice a good enough description for specifying the frequency of different possible values of: (a) the mean score of 100 tosses of a die; (b) the difference between the proportion of hearts in samples of 50 and 75 cards taken from a full pack with replacement of each card drawn before extraction of another. Similar remarks apply to the use of the sampling distributions invoked by the statistical techniques dealt with in this volume. Thus the reasons for performing a *Chi-Square Test*, i.e. a test based on a special form of Pearson's Type III, or the reasons for performing a *t-Test*, i.e. a test based on a special form of Pearson's Type VII, will depend on the character of the statistic whose sampling distribution is under consideration; but we are in a position to grasp what the reasons are only if we have some acquaintance with the algebraic properties of Pearson's Type III or Type VII, as the case may be.

For this reason, and especially because of a common pattern to which Pearson's Types conform, it will be convenient to deal with significance tests in a sequence of three chapters (14-16) rather than to set forth the rationale of a test appropriate to a particular statistical procedure in the same context as the exposition of the terms of reference of the latter. All the new tests we shall meet do in fact come within the compass of Pearson's system of which we had a preview in Chapter 6, being referable to Types I-III, VI and VII. If we recognise this at the outset, it will be possible to understand the relevance of each to the statistical procedure which invokes it without recourse to considerations derived from the geometry of hyperspace, or the use of matrix algebra. The standpoint of the author is that a statistical test based on



the assumption of a continuous score distribution is merely a device based on curve-fitting, and as such is at best an approximate description of the real world. From the standpoint of the student who is not a trained mathematician it is therefore fortunate that Pearson's system of curve-fitting by moments anticipates so many of the requirements of subsequent theory. Hence the chapters dealing with significance tests start with an elementary exposition of the properties of moments and a review of Pearson's type system in so far as it is pertinent to the end in view.

As stated, the several statistical procedures dealt with below have in common that they call for methods more elaborate than those employed in classifying and summarising quantitative data relevant to the issues dealt with in Vol. I; and it is easy for the beginner to confuse two different aims which may converge in the exposition of the algebraic rationale of any one of them. One is to specify certain relations—which we may speak of as *tautologies*—between numbers set out in a particular framework of classification, their truth having as such no necessary connexion with the theory of probability. The other is to construct summarising indices which have properties consonant with the requirements of *sampling theory*. The two aims overlap. For the interpretation of tautologies suggestive of an index whose sampling distribution is specifiable—or approximately specifiable—is essential to an understanding of the use of a statistical method, if only because its use depends on what information it summarises.

At this stage, the last remarks may not be clear to the reader as yet unaware that it is rarely profitable to read a book by starting at the beginning and continuing to the end. The author can merely hope that some will return to our last words after a first quick perusal. Here it must suffice to say that there is common ground in the task of summarising data for very different statistical techniques which employ several criteria of classification, and hence of making the logical assumptions inherent in their credentials. Consequently, we shall start (in this chapter) with the exposition of notations which have no other justification than to reduce the effort involved in recognising some purely formal relations between numerical data when assembled in a particular way. Against this background, we shall examine (Chapters 12-13) the rationale of two statistical procedures without reference to what tests we appropriately invoke when applying them.

### 11.01 THREE TYPES OF GRID

It may first be helpful if we clarify two arbitrary levels of classification, which we may distinguish as uni-dimensional and multi-dimensional. In the first category we include: (a) classifying a population (*universe* or *sample*) by stating how many or what proportion of individuals belong to a particular class distinguished by some attribute specifiable in either explicitly quantitative terms (e.g. *tall*, meaning 5 ft. 7 in. or over, or *anaemic*, meaning with an erythrocyte count of less than three million per cu. mm. of blood) or qualitative terms (e.g. *yellow*, *Protestant*, *naturalised American*); (b) classifying a population on some uniform scale as by stating how many individuals (at a given time) have a body temperature of such and such by intervals of one-tenth of a degree Fahrenheit, or an income of such and such by intervals of £50 per annum. In contra-distinction to the above the simplest sort of multi-dimensional classification (i.e. a 2-dimensional) arises when

- (i) we can assign to every individual of a population two scores (e.g. height and weight, or earned and unearned income);
- (ii) we can state how many members of each of one exclusive set of sub-populations (e.g. *Protestant*, *Catholic*, *Greek Orthodox*, *Other*) are assignable to another of a second exclusive set (e.g. *American-born*, *Naturalised*, *Other*);



- (iii) we can assign to every individual (or group) a score value (e.g. *blood calcium* level or *milk yield*) on a 1-dimensional scale, and can assign each individual (or group) to one of two or more exclusive sets of sub-populations (e.g. urban and rural, tuberculous and healthy).

The questions which prompt us to classify data in one or other way referred to in the preceding paragraph are various; but our method of assembling our data depends as much on the nature of the data as on the nature of the question. Inherent in any method we adopt are certain relations implicit in the method itself; and our preliminary task in this chapter will be to examine the problem of *summarising* data to bring into focus such relations aside from any utility they may prove to have from the standpoint of statistical theory. Corresponding to each of the 3 types of 2-dimensional classification we may thus prescribe as the first step in the summarisation of our numerical data a particular method of tabulation, i.e. a grid-wise lay-out.

For case (i) which we speak of as the *bivariate* population, it takes the form shown below as a score-frequency (more briefly, frequency) grid. The explicit ( $n_{ij}$ ) in each grid cell is the number or proportion of individuals with a unique combination of *A*-scores and *B*-scores as indicated by the entries at the head of each column and row. To each cell we can therefore assign a score function of an *A*-score ( $a_i$ ) alone (e.g.  $a_i^2$ ), of a *B*-score ( $b_j$ ) alone (e.g.  $b_j^4$ ) or of both (e.g.  $a_i b_j$ ). The reader of Volume I has made the acquaintance of this lay-out in Chapters 8-9.

<i>Border scores</i>	$a_0$	$a_1$	$a_2$	$a_3$
$b_0$	$n_{00}$	$n_{10}$	$n_{20}$	$n_{30}$
$b_1$	$n_{01}$	$n_{11}$	$n_{21}$	$n_{31}$
$b_2$	$n_{02}$	$n_{12}$	$n_{22}$	$n_{32}$

The important peculiarity of the frequency (or correlation) grid is that *each* dimension carries a set of border-scores which collectively specify the criterion of sub-classification (e.g. *height*) in that dimension, and the notation makes sufficiently explicit the corresponding frequency of the cell-score under consideration. For example, we may be interested in the distribution of score products of the form  $a_i^2 b_j^3$ . If so, we can write down the frequency of the particular product  $a_3^2 b_1^3$  as the value of  $n_{31}$  shown in the table.

The appropriate lay-out for data specified as case (ii) above is the *contingency* grid:

	<i>Protestant</i>	<i>Catholic</i>	<i>Greek</i>	<i>Other</i>	<i>TOTAL</i>
<i>American-born</i>	$n_{00}$	$n_{10}$	$n_{20}$	$n_{30}$	$N_{.0}$
<i>Naturalised</i>	$n_{01}$	$n_{11}$	$n_{21}$	$n_{31}$	$N_{.1}$
<i>Other</i>	$n_{02}$	$n_{12}$	$n_{22}$	$n_{32}$	$N_{.2}$
<i>TOTAL</i>	$N_{0.}$	$N_{1.}$	$N_{2.}$	$N_{3.}$	$N$

The totals set out here have a special importance, because it is implicit in the structure of a *true* contingency table that we can specify the number or proportion of individuals in each of the *B*-classes (here *national status*) assignable to any one of the *A*-classes (here *religious faith*). Thus the column totals ( $N_{i.}$ ), row totals ( $N_{.j}$ ) and the grand total ( $N$ ) are all fixed. It follows from this that we can fill in any cell of a row or any cell of a column, by deduction of the residual total,



if we know the entries for all the other cells of the same row or column. In a grid of  $r$  rows and  $c$  columns (i.e.  $rc$  cells excluding the column and row totals), there are thus  $c$  redundant row entries and  $r$  redundant column entries, and in all  $r + c - 1$  redundant entries, since the last cell of the last row and the last column is common to both. Hence the number of cell entries we require to know before we can complete the table is  $rc - (r + c - 1) = (r - 1)(c - 1)$ .

This restriction distinguishes what we here call a *true* contingency table from a lay-out which is superficially like it. Following Mendel one may classify pea plants exclusively by seed-coat as *yellow* or *green* and exclusively by stature as *tall* or *short*. All we may happen to know about the possible structure of a sample of  $N$  plants we may then set out in one dimension as

Tall.		Short.		Total.
Green.	Yellow.	Green.	Yellow.	
$a$	$b$	$c$	$d = (N-a-b-c)$	$N$

Alternatively, our lay-out may be

	Tall	Short	
Green	$a$	$c$	
Yellow	$b$	$d = (N-a-b-c)$	
		Total	$N$

A spurious contingency table of this sort summarises the *possibilities* rather than the actualities of an  $N$ -fold population structure; and the reader will note that we require to know  $rc - 1$  cell entries before we can complete it in the absence of the additional information our row and column totals of a true contingency table supply.

The two kinds of tabulation schematised in the foregoing remarks have this in common that the *explicit* cell entries are absolute or relative frequencies, i.e. proportions or numbers of individuals. For data specified as case (iii) above, the appropriate lay-out is a grid of which the cell entries are *scores*. If we have only one *qualitative* criterion of classification the score-grid is merely a set of scores laid out in any order within columns referable to all-or-none classes; we can set out data classifiable w.r.t. more than one qualitative criterion as below. Here the cell entries ( $x_{ij}$ ) are scores referable to individual members of the population or groups of individuals specified as members of one or other class of 2 different sets as indicated by labelling the rows and columns. Thus we may be able exclusively and simultaneously to assign the fertility rates ( $x_{ij}$ ) of groups which share the same religious faith and groups distinguished by national status, as below:

	Protestant	Catholic	Other
American-born	$x_{00}$	$x_{10}$	$x_{20}$
Naturalised	$x_{01}$	$x_{11}$	$x_{21}$
Other	$x_{02}$	$x_{12}$	$x_{22}$



The student will note that we have added no entries for score totals at the foot of the columns or margin of the rows. Nor would the addition of such information dispense with the need to make every cell entry explicit, unless we also knew how many individuals each sub-population contains. In what follows we shall explore relations between numerical characteristics of sub-populations classified in 2 or more dimensions with special reference to the correlation grid (11.02–11.04) and the score-grid (11.05).

### *Relation Between Score-Grid and Correlation Grid*

As pointed out (p. 408) in Chapter 10 and illustrated in Fig. 83 of Vol. I, it is not possible to convert a score-grid exhibiting two different criteria of classification (one referable to columns, the other to rows) into a 2-dimensional frequency grid; but it is always possible to summarise the data exhibited in a correlation grid by recourse to the alternative device of a score-grid with one explicit criterion ( $A$ , *not-A*) of classification indicated at the margin of the columns or rows. Such a score-grid of two arrays is indeed the lay-out for computation of the product-moment index as in Chapter 8 (pp. 354–355). We may in fact summarise the distribution of sixteen paired  $(x_a, x_b)$  scores in three ways as in the numerical example below:

#### (i) *Frequency Grid*

		$(x_a)$			
Border-scores		0	1	2	3
$(x_b)$	0	1	1	0	0
	1	1	3	2	0
	2	0	2	3	1
	3	0	0	1	1

#### (ii) *Bivariate Score Distribution*

$x_a \cdot x_b$	.	.	0.0	0.1	1.0	1.1	1.2	2.1	2.2	2.3	3.2	3.3
Rel. Freq.	.	.	1	1	1	3	2	2	3	1	1	1

#### (iii) *Score-Grid of 2 rows and 16 columns (one explicit criterion)*

$A$	0	0	1	1	1	1	1	1	2	2	2	2	2	2	3	3
$B$	0	1	0	1	1	1	2	2	1	1	2	2	2	3	2	3

### *Subscript Notation*

Many of the problems of manifold classification are simple, or at least amenable to simple treatment from an algebraic viewpoint. The difficulties of the beginner arise especially from the difficulty of recognising the precise meaning of the symbols. That is why we shall here use a notation which is at first sight cumbersome and to the student unaccustomed to subscript notation a little formidable. Fortunately, familiarity will breed contempt for the disinclination to get used to it. In fact, it is much easier to apply elementary mathematics, if one makes the meaning of the symbols as explicit as possible. For instance,  $b_{oh}$ ,  $b_{oa}$ ,  $b_{op}$  for British officers at home, abroad on service and abroad as prisoners of war are less confusing to work with than the  $x$ ,  $y$ ,  $z$  of the school books.



When we say that the correlation or frequency grid is the appropriate summarising lay-out for a *bivariate* population, we use population and individual respectively for an assemblage and its constituents whether human or otherwise, inanimate or living. The population, as in the model situations of Chapter 8 in Vol. I, may consist of individual human beings to each of which we can assign examination marks for two different subjects (A and B) or, as in the models of Chapter 9, a sequence of individual games each of which is specifiable in terms of the score of each of two players (A and B). For purposes of computation it may be convenient to use whole numbers as our explicit cell entries, though in fact the most suitable schema for computation by machine work will not take the grid form. For analysis of the numerical rules inherent in any such lay-out, it is more convenient to use frequencies ( $y_{ij}$ ) as defined in Chapter 3, p. 101, being the ratio of the cell number ( $n_{ij}$ ) to the grand total ( $N$ ), i.e.  $y_{ij} = (n_{ij} \div N)$ . We may then set out the grid as below :

Border scores	$a_0$	$a_1$	$a_2$	...	$a_i$	.....	TOTALS
$b_0$	$y_{00}$	$y_{10}$	$y_{20}$	...	$y_{i0}$	.....	$y_{\cdot 0}$
$b_1$	$y_{01}$	$y_{11}$	$y_{21}$	...	$y_{i1}$	.....	$y_{\cdot 1}$
$b_2$	$y_{02}$	$y_{12}$	$y_{22}$	...	$y_{i2}$	.....	$y_{\cdot 2}$
...	...	...	...	...	...	.....	...
$b_j$	$y_{0j}$	$y_{1j}$	$y_{2j}$	...	$y_{ij}$	.....	$y_{\cdot j}$
...	...	...	...	...	...	.....	...
TOTALS	$y_{0\cdot}$	$y_{1\cdot}$	$y_{2\cdot}$	...	$y_{i\cdot}$	.....	1

This schema embraces more than the structure of a bivariate population as defined above. With appropriate interpretation of  $y_{ij}$  it describes the frequency distribution of any function of the scores of two samples drawn independently from the same or identical universes. Thus the cell-score corresponding to  $y_{ij}$  is  $(a_i + b_j)$  if our concern is with the distribution of the *score-sum* and  $(a_i - b_j)$  if our concern is with the distribution of the raw-score *difference*. In any case, the column totals ( $y_{.j}$  etc.) and the row totals ( $y_{i.}$  etc.) respectively specify the frequencies of the border *A*-scores and *B*-scores, in this case the two sets of sample scores; and the product rule of *independent* assortment is equivalent to the substitution in each cell of the identity

$$y_{i..} y_{.j} = y_{ij} \quad . \quad . \quad . \quad . \quad . \quad . \quad (i)$$

In any case, for a grid of  $c$  columns and  $r$  rows labelled respectively from 0 to  $(c - 1)$  and 0 to  $(r - 1)$ , in conformity with the convention for finite difference series,

[illegible]

$$\sum_{j=0}^{(r-1)} y_{.j} = 1 = \sum_{i=0}^{(c-1)} y_{i.} \quad . \quad . \quad . \quad . \quad . \quad (iii)$$

$$\sum_{j=0}^{(r-1)} \sum_{i=0}^{(c-1)} y_{ij} = 1 = \sum_{i=0}^{(c-1)} \sum_{j=0}^{(r-1)} y_{ij}. \quad . \quad . \quad . \quad . \quad . \quad (\text{iv})$$

If  $x_{ij}$  is the cell-score of the  $i$ th column and the  $j$ th row, we may denote the mean value of the  $x$ -score for the  $i$ th column as  $M_{x.i}$ , for the  $j$ th row as  $M_{x.j}$  and for the whole grid as  $M_x$ . By definition of the mean, we then have

$$M_{x.i} = \frac{1}{y_{i.}} \sum_{j=0}^{(r-1)} y_{ij} \cdot x_{ij}; \quad M_{x.j} = \frac{1}{y_{.j}} \sum_{i=0}^{(c-1)} y_{ij} \cdot x_{ij} \quad \text{. . . . . (v)}$$

$$\sum_{i=0}^{(c-1)} y_{i.} \cdot M_{x.i} = \sum_{i=0}^{(c-1)} \sum_{j=0}^{(r-1)} y_{ij} \cdot x_{ij} = M_x = \sum_{j=0}^{(r-1)} \sum_{i=0}^{(c-1)} y_{ij} \cdot x_{ij} = \sum_{j=0}^{(r-1)} y_{.j} \cdot M_{x.j} \quad \text{(vi)}$$

*Example 1.*—From the crude data on the extreme left determine the row-means, column-means and grand means of  $x = a^2b$ .

A-scores ( $a_i$ )					Frequencies				
	0	1	2	TOTAL					TOTAL
B-scores ( $b_j$ )	1	1	3	4	8	0.025	0.075	0.1	0.2
	2	2	2	8	12	0.05	0.05	0.2	0.3
	3	3	4	5	12	0.075	0.10	0.125	0.3
	4	2	1	5	8	0.05	0.025	0.125	0.2
	TOTAL	8	10	22	40	0.2	0.25	0.55	1.0

$$x = a_i^2 b_j$$

0	1	4	...
0	2	8	...
0	3	12	...
0	4	16	...
...	...	...	...

The column means of the  $x$ -score are

$$0; \frac{(0.075 + 0.10 + 0.30 + 0.10)}{0.25}; \frac{(0.4 + 1.6 + 1.5 + 2.0)}{0.55} \\ = 0; 2.3; 10.$$

The row means are

$$\frac{(0.075 + 0.4)}{0.2}; \frac{(0.10 + 1.6)}{0.3}; \frac{(0.30 + 1.5)}{0.3}; \frac{(0.10 + 2.0)}{0.2} \\ = 2.375; 5.6; 6; 10.5.$$

The grand mean is

$$(0.075 + 0.4 + 0.10 + 1.6 + 0.30 + 1.5 + 0.10 + 2.0) = 6.075.$$

\* \* \* \* \*







$u_i = x_a^h$  throughout column  $i$  when  $h$  is an integer . . . . . (xiii)

$v_j = x_b^k$  throughout row  $j$  when  $k$  is an integer. . . . . (xiv)

$w_{ij} =$  any single-valued function of both  $x_a^h$  and  $x_b^k$  in the cell  $(i, j)$ , when  $h$  and  $k$  are integers  
or zero . . . . . (xv)

For brevity we shall also use  $M_{u \cdot j}$ ,  $M_{w \cdot j}$  for row-means,  $M_{v \cdot i}$ ,  $M_{w \cdot i}$  for column-means,  $M_u$ ,  $M_v$ ,  $M_w$  for grand means of  $u_i$ ,  $v_j$ ,  $w_{ij}$  and  $M_a$ ,  $M_b$  for the grand mean of  $x_a$  and  $x_b$ . Some of the ensuing rules apply to all the  $E$ -operations in which case we shall use  $E_0$  generically.

### Rule 1. Redundant Operations

Since  $u_i$  is constant throughout the column and  $v_j$  throughout the row

$$\begin{aligned} E_{b \cdot a}(u_i) &= u_i; E_{a \cdot b}(v_j) = v_j, \\ \therefore E(u_i) &= E_a \cdot E_{b \cdot a}(u_i) = E_a(u_i); E(v_j) = E_b \cdot E_{a \cdot b}(v_j) = E_b(v_j) \end{aligned} \quad (xvi)$$

Notice, however, that

$$E_{b \cdot a}(v_j) = M_{v \cdot i} \text{ and } E_a(M_{v \cdot i}) = M_v; E_{a \cdot b}(u_i) = M_{u \cdot j} \text{ and } E_b(M_{u \cdot j}) = M_u \quad (xvii)$$

### Rule 2. Change of Origin and Scale

From elementary arithmetical considerations we know that the effect of a scalar constant is multiplicative and that of an additive constant is additive in the process of extracting the mean value of a score function, i.e.

$$E_0(K \cdot u_{ij} + C) = K \cdot E_0(u_{ij}) + C \quad (xviii)$$

Since  $u_i$  is a constant in the B-dimension and  $v_j$  in the A-dimension of the grid, we have

$$\begin{aligned} E(u_i v_j) &= E_a \cdot E_{b \cdot a}(u_i v_j) = E_a[u_i \cdot E_{b \cdot a}(v_j)]; \\ E(u_i v_j) &= E_b \cdot E_{a \cdot b}(u_i v_j) = E_b[v_j \cdot E_{a \cdot b}(u_i)]. \end{aligned}$$

We may write the above as

$$E_a(u_i \cdot M_{v \cdot i}) = E(u_i \cdot v_j) = E_b(v_j \cdot M_{u \cdot j}) \quad (xix)$$

*Example 2.*—Put  $y_a = (3x_a + 1)$  and  $z_b = (2x_b + 5)$  in the foregoing numerical illustration (Ex. 11.01) so that the column border-scores become 1, 4, 7, 10 and the row border-scores 5, 7, 9, 11. Now find the row-means, column-means and grand mean of  $y_a^3$ ,  $z_b^2$ . Compute also the corresponding mean values  $x_a^3$  and  $x_b^2$ , and check (xix) when  $u_i = x_a^2$  and  $v_j = x_b$ .

### Rule 3. Sum or Difference of Means

It follows from the last two rules that

$$\begin{aligned} E_{b \cdot a}(u_i \pm v_j) &= u_i \pm E_{b \cdot a}(v_j) = u_i \pm M_{v \cdot i}; \\ E_{a \cdot b}(u_i \pm v_j) &= E_{a \cdot b}(u_i) \pm v_j = M_{u \cdot j} \pm v_j, \\ \therefore E(u_i \pm v_j) &= E_a \cdot E_{b \cdot a}(u_i \pm v_j) = E_a(u_i) \pm E_a(M_{v \cdot i}), \\ \therefore E(u_i \pm v_j) &= E(u_i) \pm E(v_j) \end{aligned} \quad (xx)$$

If  $s_{ij} = (u_i + v_j)$ , or  $d_{ij} = (u_i - v_j)$  we may write this as

$$M_s = M_u + M_v; M_d = M_u - M_v.$$

In connexion with moments, it is important to recall the rule in this form since

$$s_{ij} - M_s = (u_i - M_u) + (v_j - M_v) \quad (xxi)$$

$$d_{ij} - M_d = (u_i - M_u) - (v_j - M_v) \quad (xxii)$$

*Example 3.*—Test this rule by making a sum and difference table from the data of Example 1 for  $w_{ij} = (u_i \pm v_j)$  when  $u_i = x_a^3$  and  $v_j = x_b^2$  as below :

Sum ( $x_a^3 + x_b^2$ )			Frequencies ( $\times 40$ )			Differences ( $x_a^3 - x_b^2$ )		
1	2	9	1	3	4	— 1	0	7
4	5	12	2	2	8	— 4	— 3	4
9	10	17	3	4	5	— 9	— 8	— 1
16	17	24	2	1	5	— 16	— 15	— 8

#### Rule 4. Partition of Variance

One of the most important summarising tautologies of the grid concerns the partition of variance w.r.t. either set of border-scores  $x_a$  and  $x_b$ . We shall write for the total variance

$$V_a = E(x_a - M_a)^2 = E(x_a^2) - M_a^2 \quad \text{and} \quad V_b = E(x_b - M_b)^2 = E(x_b^2) - M_b^2 \quad . \quad (\text{xxiii})$$

For the variance within the row or within the column we have

$$V_{a \cdot b} = E_{a \cdot b}(x_a - M_{a \cdot b})^2 = E_{a \cdot b}(x_a^2) - M_{a \cdot b}^2$$

and

$$V_{b \cdot a} = E_{b \cdot a}(x_b - M_{b \cdot a})^2 = E_{b \cdot a}(x_b^2) - M_{b \cdot a}^2$$

The mean values of the above are

$$M(V_{a \cdot b}) = E(x_a^2) - E_b(M_{a \cdot b}^2) \quad \text{and} \quad M(V_{b \cdot a}) = E(x_b^2) - E_a(M_{b \cdot a}^2) \quad . \quad (\text{xxiv})$$

For the variance of the row- and column-means we have

$$V(M_{a \cdot b}) = E_b(M_{a \cdot b} - M_a)^2 = E_b(M_{a \cdot b}^2) - M_a^2$$

and

$$V(M_{b \cdot a}) = E_a(M_{b \cdot a} - M_b)^2 = E_a(M_{b \cdot a}^2) - M_b^2 \quad . \quad . \quad . \quad (\text{xxv})$$

By combining (xxiii)–(xxv) we have

$$V_a = M(V_{a \cdot b}) + V(M_{a \cdot b}) \quad \text{and} \quad V_b = M(V_{b \cdot a}) + V(M_{b \cdot a}) \quad . \quad . \quad (\text{xxvi})$$

The reader should note that this is a necessary numerical property of *any* set of numbers laid out as a score-frequency grid ; and should test it by recourse to the data of Example 1.

#### Rule 5. Moments of a Discrete Distribution

With the aid of Rules 1-3 we may compress into a single expression several important properties of moments. As defined in Chapter 6 of Vol. I, the  $p$ th zero moment ( $\mu_p$ ) of a score distribution is the mean value of the  $p$ th power of the scores and the  $p$ th mean moment ( $m_p$ ) is the mean value of the  $p$ th power of the score deviations from their mean value. For the column border-scores ( $x_a$ ) we thus write

$${}_a\mu_p = E(x_a^p) \quad \text{and} \quad {}_a m_p = E(x_a - M_a)^p = E(x_a - {}_a\mu_1)^p.$$

The properties we shall now exhibit are derivable from the mean value of a single expression involving both sets of border-scores. On the assumption that  $p$  is an integer, we define it as

$$[(x_a - C) \pm (x_b - K)]^p = F = [(x_a \pm x_b) - (C \pm K)]^p \quad . \quad . \quad (\text{xxvii})$$









*Example 5.*—Test (xxxvi) for  $h = 3$  and  $k = 2$  from the data of Example 1.

The following identity will sometimes prove useful. If  $X_a = (x_a - M_a)$  and  $X_b = (x_b - M_b)$ :

$$\begin{aligned} x_b \cdot X_a &= x_a \cdot x_b - M_a \cdot x_b \text{ and } x_a \cdot X_b = x_a \cdot x_b - M_b \cdot x_a, \\ \therefore E(x_b \cdot X_a) &= E(x_a \cdot x_b) - M_a \cdot E(x_b) \text{ and } E(x_a \cdot X_b) = E(x_a \cdot x_b) - M_b \cdot E(x_a), \\ \therefore E(x_b \cdot X_a) &= E(x_a \cdot x_b) - M_a \cdot M_b = E(x_a \cdot X_b), \\ \therefore E(x_b \cdot X_a) &= \text{Cov}(x_a, x_b) = E(x_a \cdot X_b) \quad \dots \dots \dots \quad \text{(xxxix)} \end{aligned}$$

We may likewise express the covariance in the following way

$$\begin{aligned} E(x_a - M_a)(x_b - M_b) &= E_a \cdot E_b \cdot a \cdot X_a(x_b - M_b) = E_a \cdot X_a[E_b \cdot a(x_b - M_b)], \\ \therefore E(x_a - M_a)(x_b - M_b) &= E_a \cdot X_a(M_b \cdot a - M_b). \end{aligned}$$

Similarly,

$$\begin{aligned} E(x_a - M_a)(x_b - M_b) &= E_b \cdot E_a \cdot b \cdot X_b(x_a - M_a) = E_b \cdot X_b[E_a \cdot b(x_a - M_a)], \\ \therefore E(x_a - M_a)(x_b - M_b) &= E_b \cdot X_b(M_a \cdot b - M_a). \end{aligned}$$

Whence we have

$$E_b(x_b - M_b)(M_a \cdot b - M_a) = \text{Cov}(x_a, x_b) = E_a(x_a - M_a)(M_b \cdot a - M_b) \quad \dots \quad \text{(xl)}$$

### 11.03 THE INDEPENDENCE GRID

To say that two distributions such as those of the border-scores  $x_a, x_b$  of our grid in 11.02 are *independent* in the statistical sense of the term is to say that the cell frequencies are in accord with the principle of equipartition of opportunity for association, i.e. with the product rule specified by (i) of 11.02, viz.:

$$y_{ij} = y_{i \cdot} \cdot y_{\cdot j}.$$

If we now go back to our code, we see that (vii) and (viii) of 11.02 then mean

$$\begin{aligned} E_b \cdot a(\cdot \cdot \cdot) &\equiv \sum_{j=0}^{(r-1)} y_{\cdot j}(\cdot \cdot \cdot) \equiv E_b(\cdot \cdot \cdot); \\ E_a \cdot b(\cdot \cdot \cdot) &\equiv \sum_{i=0}^{(c-1)} y_{i \cdot}(\cdot \cdot \cdot) \equiv E_a(\cdot \cdot \cdot). \end{aligned}$$

If, as before,  $u_i = x_a^h$  and  $v_j = x_b^k$ , this means

$$M_{v \cdot i} = M_v = E_b(v_j) \quad \text{and} \quad M_{u \cdot j} = M_u = E_a(u_i).$$

In the notation of moments, we write this as

$$b \cdot a \mu_k = E_b \cdot a(x_b^k) = b \mu_k \quad \text{and} \quad a \cdot b \mu_h = E_a \cdot b(x_a^h) = a \mu_h \quad \dots \quad \text{(i)}$$

In the same way, we interpret (xix) in 11.02 as follows:

$$\begin{aligned} E_a(u_i \cdot M_{v \cdot i}) &= E(u_i \cdot v_j) = E_b(v_j \cdot M_{u \cdot j}), \\ \therefore M_v \cdot E_a(u_i) &= E(u_i \cdot v_j) = M_u \cdot E_b(v_j), \\ \therefore E(u_i \cdot v_j) &= M_u \cdot M_v. \end{aligned}$$

In the same notation of (xxxviii) this is equivalent to

$$\mu_{hk} = \mu_h \cdot \mu_k \quad \dots \quad \text{(ii)}$$

Whence independence implies in virtue of (xxxvi):

$$\text{Cov}(x_a^h, x_b^k) = E(x_a^h - a \mu_h)(x_b^k - b \mu_k) = 0 \quad \dots \quad \text{(iii)}$$

When  $h = 1 = k$  we write this as  $\text{Cov}(x_a, x_b) = 0$ .

$$d\mu_p = \sum_{z=0}^{z=p} (-1)^z p_{(z)} d\mu_z \cdot b\mu_{p-z} . \quad . \quad . \quad . \quad . \quad . \quad . \quad (xi)$$



For example, we have

$$\begin{aligned}s\mu_3 &= {}_a\mu_3 + 3{}_a\mu_2 \cdot {}_b\mu_1 + 3{}_a\mu_1 \cdot {}_b\mu_2 + {}_b\mu_3; \\ {}_a\mu_3 &= {}_a\mu_3 - 3{}_a\mu_2 \cdot {}_b\mu_1 + 3{}_a\mu_1 \cdot {}_b\mu_2 - {}_b\mu_3.\end{aligned}$$

If we recall that  $m_1 = 0$  for any distribution, we may likewise write

$${}_sm_3 = {}_am_3 + {}_bm_3 \text{ and } {}_am_3 = {}_am_3 - {}_bm_3.$$

We shall use the foregoing results extensively in Chapter 14.

\* \* \* \* \*

Before proceeding, the student may with profit perform the following exercises by recourse to the data of the following table in which the cell entries are whole numbers :

Border-scores	0	1	2	3
2	6	2	8	4
4	15	5	20	10
6	9	3	12	6

- (i) Satisfy yourself that the cell entries obey the product rule.
- (ii) Find the third zero and fourth mean moments of the  $A$ -scores in each of the rows and the variance and fifth zero moment of the  $B$ -scores in each of the columns.
- (iii) Verify equation (iii) above for  $h = 1 = k$ , and for  $h = 3, k = 5$ .
- (iv) Make tables of the score-sum and the raw-score difference distributions and verify (viii)–(xi) for  $p = 4$ .

\* \* \* \* \*

As a particular case of (i), we have in (xxv) of 11.02

$$M_{b \cdot a} = M_b = {}_b\mu_1 \text{ and } M_{a \cdot b} = M_a = {}_a\mu_1.$$

Thus we have

$$V(M_{b \cdot a}) = E_a(M_{b \cdot a} - M_b)^2 = 0 = E_b(M_{a \cdot b} - M_a)^2 = V(M_{a \cdot b}) \quad . \quad . \quad (xii)$$

We now recall the definition of the *correlation ratios* defined in Chapter 9 of Vol. I, viz. :

$$\eta_{ab}^2 = \frac{V(M_{a \cdot b})}{V_a} \text{ and } \eta_{ba}^2 = \frac{V(M_{b \cdot a})}{V_b}.$$

Thus the correlation ratios of a bivariate distribution are both zero if the component distributions are statistically independent. The product-moment correlation coefficient is necessarily so by definition, since  $Cov(x_a, x_b) = 0$ . While this is however a necessary consequence of independence, it is important to appreciate that it is not a sufficient criterion of independence. That is to say, a joint distribution may have zero covariance when the product rule does not apply. In 11.08 we shall look at a few model situations which illustrate this possibility.

#### 11.04 TAUTOLOGIES OF CORRELATION

We are now in a position to set out certain summarising tautologies which we shall make use of in Chapters 12 and 20. We say that there is perfect correlation between two sets of scores if

$$Cov(x_a, x_b) = k_{ab} \cdot V_b \quad . \quad . \quad . \quad . \quad . \quad (v)$$





We shall now write for the expression on the left

$$D = (M_{b \cdot a} - M_b) - \frac{C_{ab}}{V_a}(x_a - M_a).$$

Thus linear regression of the  $B$ - on the  $A$ -score signifies that  $D = 0$ , as must be true if both its mean value and its variance ( $V_d$ ) are each zero. That the mean value of  $D$  is zero is evident, since

$$E_a(M_{b \cdot a} - M_b) = E_a(M_{b \cdot a}) - M_b = M_b - M_b,$$

and since  $C_{ab}$  and  $V_a$  are constants and  $E_a(x_a - M_a) = 0$ , the second term of the expression vanishes.

$$\begin{aligned} V_d &= E_a(D^2) - [E(D)]^2 = E_a(D^2) \\ &= E_a\left\{(M_{b \cdot a} - M_b)^2 - \frac{2C_{ab}(M_{b \cdot a} - M_b)(x_a - M_a)}{V_a} + \frac{C_{ab}^2}{V_a^2}(x_a - M_a)^2\right\} \\ &= E_a(M_{b \cdot a} - M_b)^2 - \frac{2C_{ab}}{V_a}E_a(M_{b \cdot a} - M_b)(x_a - M_a) + \frac{C_{ab}^2}{V_a^2}E_a(x_a - M_a)^2. \end{aligned}$$

In the above  $E_a(x_a - M_a)^2 = V_a$ ,  $E_a(M_{b \cdot a} - M_b)^2 = V(M_{b \cdot a})$ ; and we may write as for (xl) in 11.02:

$$\begin{aligned} E_a(M_{b \cdot a} - M_b)(x_a - M_a) &= E_a(x_a \cdot M_{b \cdot a}) - M_a \cdot E_a(M_{b \cdot a}) - M_b \cdot E_a(x_a) + M_a M_b \\ &= E(x_a \cdot x_b) - M_a M_b = \text{Cov}(x_a, x_b), \\ \therefore V_d &= V(M_{b \cdot a}) - \frac{C_{ab}^2}{V_a}. \end{aligned}$$

When  $D = 0$ , so that  $V_d = 0$ , as must be true when there is linear regression of the  $B$ - on the  $A$ -score:

$$\begin{aligned} V(M_{b \cdot a}) &= \frac{C_{ab}^2}{V_a}, \\ \therefore \frac{V(M_{b \cdot a})}{V_b} &= \frac{C_{ab}^2}{V_a \cdot V_b}, \\ \therefore \eta_{ba}^2 &= r_{ab}^2 \end{aligned}$$

The sufficient, as well as the necessary condition for linear regression of the  $A$ - on the  $B$ -score is deducible in the same way. In virtue of (i) linear regression in either dimension implies that for *perfect* correlation

$$\eta_{ab}^2 = 1 = r_{ab}^2.$$

Thus the linear regression in *either* dimension of the grid guarantees that  $r_{ab}$  should have its essential summarising property, namely limits of  $\pm 1$ .

\* \* \* \* \*

In practice it will rarely, if ever, happen that regression is strictly linear; and it will be our concern in Chapter 18 to examine situations in which the composition of a sample is consistent with the assumption that regression is linear in the parent bivariate universe, though the relation defined by (ii) above is not exactly true of the particular set of data. We are still free to define a constant  $k_{ba}$  in terms of sample covariance and sample variances in accordance with (iv) and



$$x_{r,a} - M_b = k_{ba}(x_a - M_a) \quad . \quad . \quad . \quad . \quad (\text{xiv})$$
$$\begin{aligned} E_a(x_{r.a}) - M_b &= k_{ba}.E_a(x_a - M_a) = 0, \\ \therefore E_a(x_{r.a}) &= M_b \quad . \quad . \quad . \quad . \quad . \quad . \quad (\text{xv}) \end{aligned}$$
$$k_{ba} = \frac{\text{Cov}(x_a, x_b)}{V_a} = r_{ab} \frac{\sigma_b}{\sigma_a}.$$

Within the sample column, i.e. for a fixed value of  $A$ ,

$$\begin{aligned} E_{b \cdot a}(x_{r \cdot a}) &= x_{r \cdot a}, \\ \therefore E_{b \cdot a}(x_{b \cdot a} \cdot x_{r \cdot a}) &= M_{b \cdot a} \cdot x_{r \cdot a}, \\ \therefore E(x_{b \cdot a} \cdot x_{r \cdot a}) &= E_a(M_{b \cdot a} \cdot x_{r \cdot a}) \quad . \quad . \quad . \quad (\text{xvi}) \end{aligned}$$

[illegible]
$$E(x_{r.a} - M_b)^2 = k_{ba}^2 \cdot E(X_a^2) = r_{ab}^2 \cdot V_b . \quad (\text{xviii})$$

- (i)  $B$ -scores from corresponding values of  $x_{r..a}$ , i.e.  $E(x_{b..a} - x_{r..a})^2$ ;
- (ii) Mean  $B$ -scores from same, i.e.  $E(M_{b..a} - x_{r..a})^2$ .

$$E(x_{b..a} - M_{b..a})^2 = M(V_{b..a}) = (1 - \eta_{ba}^2)V_b \quad . \quad . \quad . \quad (\text{xix})$$
$$\begin{aligned} E(x_{b.a} - x_{r.a})^2 &= E(\overbrace{x_{b.a} - M_b}^{\text{residual}} - \overbrace{x_{r.a} - M_b})^2 \\ &= E(x_{b.a} - M_b)^2 + E(x_{r.a} - M_b)^2 - 2E(x_{b.a} - M_b)(x_{r.a} - M_b). \end{aligned}$$





On this understanding, if  $b_{j,i}$  denotes any  $B$ -score value in the  $i$ th column:

$$E(x_{b..a} - M_{b..a})^2 = \frac{1}{n} \sum_{i=1}^{i=c} r_i \sum_{j=1}^{j=r_i} (b_{j..i} - M_{b..i})^2 = \frac{1}{n} S_{bm};$$

$$E(M_{b..a} - x_{r..a})^2 = \frac{1}{n} \sum_{i=1}^c r_i (M_{b..i} - x_{r..i})^2 = \frac{1}{n} S_{mr};$$

$$E(x_{r.a} - M_b)^2 = \frac{1}{n} \sum_{i=1}^{i=c} r_i (x_{r.i} - M_b)^2 = \frac{1}{n} S_{rb}.$$

We then have

$$n(1 - \eta_{ba}^2)V_b = S_{bm} = \sum_{i=1}^{i=c} \sum_{j=1}^{j=r_i} r_i(b_{j \cdot i} - M_{b \cdot i})^2 \quad . \quad . \quad . \quad (\text{xxviii})$$

$$n(\eta_{ba}^2 - r_{ab}^2)V_b = S_{mr} = \sum_{i=1}^{i=c} r_i(M_{b.i} - x_{r.i})^2 \quad . \quad . \quad . \quad . \quad . \quad (\text{xxix})$$

$$n \cdot r_{ab}^2 \cdot V_b = S_{rb} = \sum_{i=1}^{i=c} r_i(x_{r.i} - M_b)^2 \quad . \quad . \quad . \quad . \quad . \quad (\text{xxx})$$

There still remain two important tautologies of a correlation grid for future reference. Let us denote by  $r_{bm}$  the correlation coefficient between the  $B$ -score distribution and the  $B$ -score column means. This merely signifies that we replace the border  $A$ -scores by corresponding values of  $M_{b..a}$  as written at the foot in most of the grids of Chapter 9 in Vol. I. To say that regression is exactly linear is to say that the substitution merely involves a change of scale and/or origin; but we have seen in Chapter 8 (p. 353) that change of scale and origin of either set of scores does not affect the value of  $r$ . Thus linear regression implies the identity

$$r_{hm} = r_{gh}.$$

This is implicit in other identities cited above. By definition

$$r_{bm} = \frac{Cov(M_{b \cdot a}, x_b)}{\sqrt{V(M_{b \cdot a}) \cdot V_b}}$$

In the above, linear regression signifies that

$$\begin{aligned} Cov(M_{b \cdot a}, x_b) &= E(M_{b \cdot a} - M_b)X_b = k_{ba} \cdot E(X_a \cdot X_b) \\ &= k_{ba} Cov(x_a, x_b). \end{aligned}$$

Whence, from (iv) above,

$$\text{Cov}(M_{b \cdot a}, x_b) = k_{ba}^2 \cdot V_a.$$

Also from (xi) above, when regression is linear,

$$\sqrt{V(M_{b \cdot a})V_b} = \sqrt{k_{ba}^2 \cdot V_a V_b} = k_{ba}\sigma_a\sigma_b.$$

Thus we have

$$r_{bm} = k_{ba} \frac{\sigma_a}{\sigma_b} = r_{ab}. \quad . \quad . \quad . \quad . \quad . \quad (\text{xxxix})$$

The reader will later find that this result is important in connexion with the definition of the multiple regression coefficient (Chapter 18).

## EXERCISE 11.04

The student may check numerically any of the relations of this section by recourse to the models cited as Examples and Exercises in Chapter 9 of Vol. I or by reference to the Model of Fig. 90.

The reader will also find ample opportunities for testing the tautologies involving relations between  $B$ -scores, mean  $B$ -scores for a fixed  $A$ -score, and hypothetical regression scores at the end of this section by interchanging members of any number of consecutive pairs of columns in the Model of Fig. 93. This will have the effect of displacing individual mean column values from what we shall later define as the line of best fit through the whole assemblage of mean column scores.

## 11.05 TAUTOLOGIES OF THE SCORE-GRID

In Chapter 10 of Vol. I we have explored what assumptions we make about the structure of a universe, when we attempt to make a balance sheet with respect to sources of variation from the information a sample supplies; but we did not attempt to specify how we estimate the components of variation from sample data from the universe of our Handicap Score-grid Model. When we later seek a rationale for the statistical procedures commonly subsumed by the expression *analysis of variance*, it will be helpful to be clear about what characteristics of a score-grid are *tautologies of any such lay-out of numbers*, regardless of considerations relevant to statistical principles or of what conclusions are our preoccupation when concerned with sources of variation.

The notation employed in the preceding sections refers only to the type of grid which exhibits cell frequencies referable to cell-scores themselves functionally related to either or both of two sets of border-scores. When discussing the type elsewhere spoken of as a *score-grid*, it is necessary to modify our symbolism. In such a 2-dimensional lay-out (p. 429) of  $c$  columns and  $r$  rows, each cell entry is a single score with gross frequency  $(cr)^{-1}$ . Its frequency within a column is  $r^{-1}$ , and within a row  $c^{-1}$ . If  $x_{ij}$  is the cell-score of column  $i$  and row  $j$ , we may define below

TABLE 1

	Mean	Variance
Whole grid . . . .	$M = \frac{1}{rc} \sum_{i=1}^c \sum_{j=1}^r x_{ij}$	$\frac{1}{rc} \sum_{i=1}^c \sum_{j=1}^r (x_{ij} - M)^2 = V$
Within-row ( $j$ th) scores . .	$M_j = \frac{1}{c} \sum_{i=1}^c x_{ij}$	$\frac{1}{c} \sum_{i=1}^c (x_{ij} - M_j)^2 = V_j$
Within-column ( $i$ th) scores . .	$M_i = \frac{1}{r} \sum_{j=1}^r x_{ij}$	$\frac{1}{r} \sum_{j=1}^r (x_{ij} - M_i)^2 = V_i$
Row means . . . .	$M = \frac{1}{r} \sum_{j=1}^r M_j$	$\frac{1}{r} \sum_{j=1}^r (M_j - M)^2 = V(M_r)$
Column means . . . .	$M = \frac{1}{c} \sum_{i=1}^c M_i$	$\frac{1}{c} \sum_{i=1}^c (M_i - M)^2 = V(M_c)$



TABLE 2

	Mean	Variance
Whole grid . . .	$E_c \cdot E_r(x_{rc}) = M$	$E_c \cdot E_r(x_{rc} - M)^2 = V = E_c \cdot E_r(x_{rc}^2) - M^2$
Within-row scores . .	$E_c(x_{rc}) = M_r$	$E_c(x_{rc} - M_r)^2 = V_r = E_c(x_{rc}^2) - M_r^2$
Within-column scores . .	$E_r(x_{rc}) = M_c$	$E_r(x_{rc} - M_c)^2 = V_c = E_r(x_{rc}^2) - M_c^2$
Row means . . .	$E_r(M_r) = M$	$E_r(M_r - M)^2 = V(M_r) = E_r(M_r^2) - M^2$
Column means . . .	$E_c(M_c) = M$	$E_c(M_c - M)^2 = V(M_c) = E_c(M_c^2) - M^2$

$$\begin{aligned} \frac{1}{rc} \sum_{j=1}^{j=r} \sum_{i=1}^{i=c} (\dots) &\equiv E(\dots) \equiv \frac{1}{cr} \sum_{i=1}^{i=c} \sum_{j=1}^{j=r} (\dots), \\ \frac{1}{r} \sum_{j=1}^{j=r} (\dots) &\equiv E_r \quad \text{and} \quad \frac{1}{c} \sum_{i=1}^{i=c} (\dots) \equiv E_c, \\ \therefore E_r \cdot E_c(\dots) &\equiv E(\dots) \equiv E_c \cdot E_r(\dots). \end{aligned} \tag{i}$$

$$E_r(V_r) = M(V_r) \quad \text{and} \quad E_c(V_c) = M(V_c).$$

$$E_r(A.u_{rc} + k) = A.E_r(u_{rc}) + k \quad \text{and} \quad E_c(A.u_{rc} + k) = A.E_c(u_{rc}) + k :$$

$$E_r(A.P_r + k) = A.E_r(P_r) + k \quad \text{and} \quad E_c(A.P_c + k) = A.E_c(P_c) + k.$$

Since  $M_c$  and  $V_c$  are constants w.r.t. rows within a column as are  $M_r$  and  $V_r$  w.r.t. columns within a row,

$$E_r(M_c) = M_c \quad \text{and} \quad E_r(V_c) = V_c;$$

$$E_c(M_r) = M_r \quad \text{and} \quad E_c(V_r) = V_r.$$

The following identities implicit in our definitions will be useful in what follows :

$$E(M_c \cdot x) = E_c[M_c \cdot E_r(x_{rc})] = E_c(M_c^2) \quad . \quad . \quad . \quad . \quad . \quad . \quad (\text{ii})$$

$$E(M_r \cdot x) = E_r[M_r \cdot E_c(x_{rc})] = E_r(M_r^2) \quad . \quad . \quad . \quad . \quad . \quad . \quad (iii)$$

$$E(M.x) = M.E(x_{rc}) = M^2 \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (\text{iv})$$

$$E(M_r, M_c) = E_r[M_r, E_c(M_c)] = E_r(M_r, M) = M^2 \quad . \quad . \quad . \quad (v)$$

[illegible]







## OPERATIONS IN A 3-DIMENSIONAL SCORE-GRID

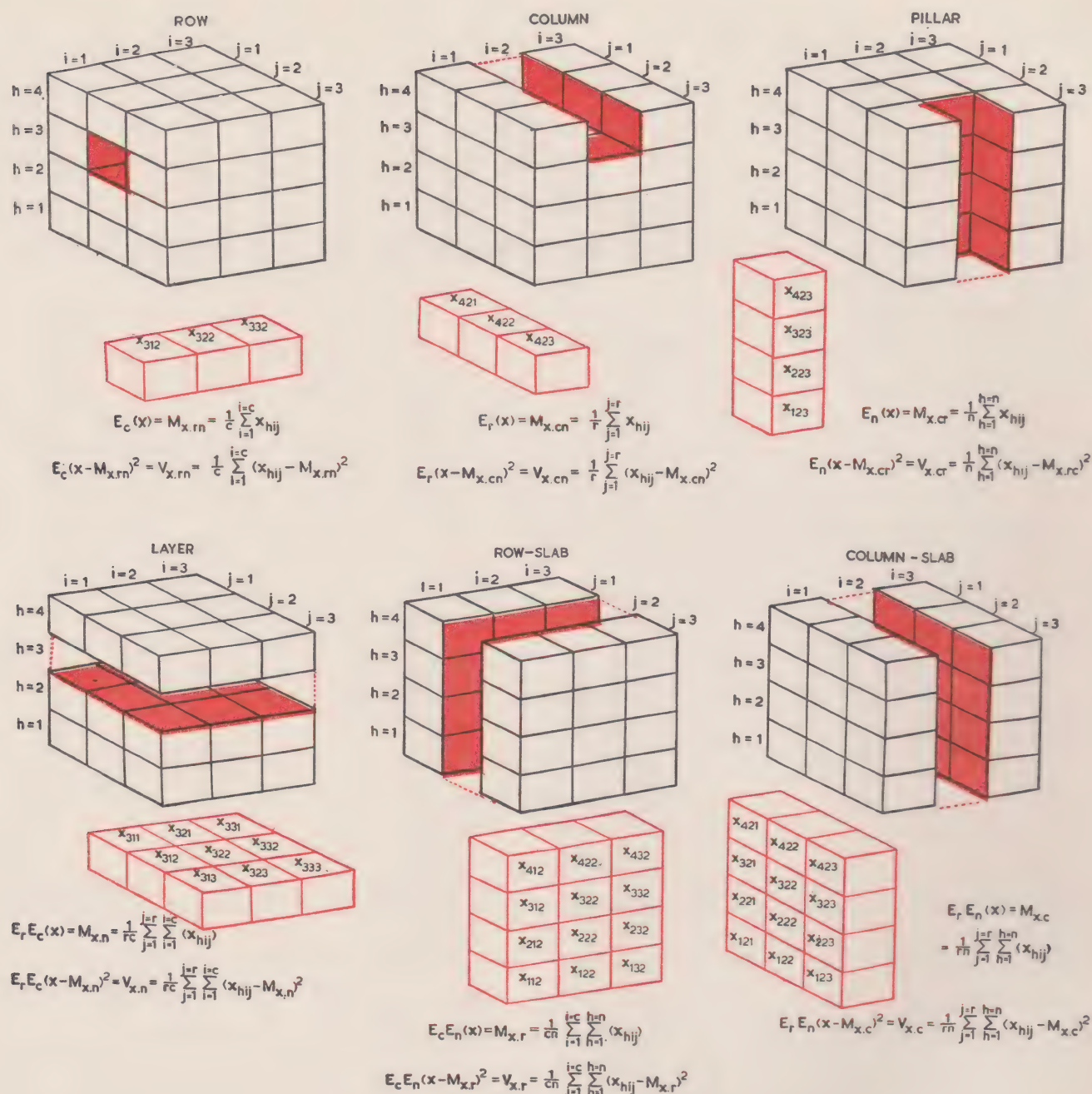


FIG. 85. The 3-dimensional Score-Grid.

The foregoing relations refer to a 2-dimensional grid admitting 2 criteria of classifying the constituent scores. Let us now accommodate 3 class specifications in a grid (Fig. 85) we can visualise by addition of a vertical dimension of layers we label as  $h = 1, 2, 3 \dots n$ . We shall denote our scores as  $x_{hij}$  accordingly. The total number of cells in the grid is then  $nrc$  distributed as follows :

<i>Layers</i>	<i>rc</i> cells	<i>Row-Slabs</i>	<i>nc</i> cells	<i>Column-Slabs</i>	<i>nr</i> cells
<i>Pillars</i>	<i>n</i> cells	<i>Rows</i>	<i>c</i> cells	<i>Columns</i>	<i>r</i> cells



$$E_n(\dots) \equiv \frac{1}{n} \sum_{h=1}^{h=n} (\dots). \quad (\text{xv})$$
$$E_r . E_c . E_n \equiv E_c . E_r . E_n \text{ etc.,}$$
$$E_c \cdot E_r(V_{x.rc}) = M(V_{x.rc}); E_r(V_{x.r}) = M(V_r); E_c(V_{x.c}) = M(V_{x.c}).$$

*Within layers :*

$$E_c(V_{x.cn}) + E_c(M_{x.cn} - M_{x.n})^2 = V_{x.n} = E_r(V_{x.rn}) + E_r(M_{x.rn} - M_{x.n})^2 \quad . \quad (\text{xvi})$$

*Within row-slabs :*

$$E_n(V_{x..rn}) + E_n(M_{x..rn} - M_{x..r})^2 = V_{x..r} = E_c(V_{x..rc}) + E_c(M_{x..rc} - M_{x..r})^2 \quad (\text{xvii})$$

*Within column-slabs :*

$$E_n(V_{x.cn}) + E_n(M_{x.cn} - M_{x.c})^2 = V_{x.c} = E_r(V_{x.rc}) + E_r(M_{x.rc} - M_{x.c})^2 \quad .(xviii)$$

$$E_r(M_{x \cdot r} - M_x)^2 + E_r(V_{x \cdot r}) = V_x.$$
$$E_c(M_{x,c} - M_x)^2 + E_c(V_{x,c}) = V_x.$$
$$E_n(M_{x \cdot n} - M_x)^2 + E_n(V_{x \cdot n}) = V_x.$$
$$\begin{aligned} E_r \cdot E_c(M_{x.rc} - M_x)^2 &= E_r \cdot E_c(M_{x.rc}^2) - 2M_x \cdot E_r \cdot E_c(M_{x.rc}) + M_x^2 \\ &= E_r \cdot E_c(M_{x.rc}^2) - M_x^2; \end{aligned}$$

$$E_r \cdot E_c(V_{x \cdot rc}) = E_r \cdot E_c \cdot E_n(x_{ncr}^2) - E_r \cdot E_c(M_{x \cdot rc}^2),$$

$$\therefore E_r \cdot E_c (M_{x.rc} - M_x)^2 + E_r \cdot E_c (V_{x.rc}) = E(x_{ncr}^2) - M_x^2 = V_x.$$

Similarly we may derive and write without ambiguity, the two following identities :

$$V(M_{x.cn}) + M(V_{x.cn}) = V_x = V(M_{x.rn}) + M(V_{x.rn}).$$

TABLE 3

	Means	Variance of cell scores <i>within</i>	Variance of Means
Pillar	$M_{x.er} = E_n(x_{h.er})$	$E_n(x_{h.er}^2) - M_{x.er}^2 = V_{x.er}$ $= E_n(x_{h.er}^2) - M_{x.er}^2$	$E_r.E_n(M_{x.er}^2 - M_x^2) = V(M_{x.er})$ $= E_r.E_n(M_{x.er}^2) - M_x^2$
Row	$M_{x.rn} = E_c(x_{i.rn})$	$E_c(x_{i.rn}^2) - M_{x.rn}^2 = V_{x.rn}$ $= E_c(x_{i.rn}^2) - M_{x.rn}^2$	$E_r.E_n(M_{x.rn}^2 - M_x^2) = V(M_{x.rn})$ $= E_r.E_n(M_{x.rn}^2) - M_x^2$
Column	$M_{x.cn} = E_r(x_{j.cn})$	$E_r(x_{j.cn}^2) - M_{x.cn}^2 = V_{x.cn}$ $= E_r(x_{j.cn}^2) - M_{x.cn}^2$	$E_n.E_c(M_{x.cn}^2 - M_x^2) = V(M_{x.cn})$ $= E_n.E_c(M_{x.cn}^2) - M_x^2$
Layer	$M_{x.n} = E_r.E_c(x_{ij.n})$ ( $= E_r.M_{x.nr} = E_c.M_{x.nc}$ )	$E_r.E_c(x_{ij.n}^2) - M_{x.n}^2 = V_{x.n}$ $= E_r.E_c(x_{ij.n}^2) - M_{x.n}^2$	$E_n(M_{x.n}^2 - M_x^2) = V(M_{x.n})$ $= E_n(M_{x.n}^2) - M_x^2$
Row slab	$M_{x.r} = E_c.E_n(x_{hi.r})$ ( $= E_n.M_{x.nr} = E_c.M_{x.er}$ )	$E_c.E_n(x_{hi.r}^2) - M_{x.r}^2 = V_{x.r}$ $= E_c.E_n(x_{hi.r}^2) - M_{x.r}^2$	$E_r(M_{x.r}^2 - M_x^2) = V(M_{x.r})$ $= E_r(M_{x.r}^2) - M_x^2$
Column slab	$M_{x.c} = E_r.E_n(x_{hj.c})$ ( $= E_n.M_{x.nc} = E_r.M_{x.er}$ )	$E_r.E_n(x_{hj.c}^2) - M_{x.c}^2 = V_{x.c}$ $= E_r.E_n(x_{hj.c}^2) - M_{x.c}^2$	$E_c(M_{x.c}^2 - M_x^2) = V(M_{x.c})$ $= E_c(M_{x.c}^2) - M_x^2$
Whole grid	$M_x = E_r.E_c.E_n(x_{hij})$ ( $= E_c.M_{x.nc} = E_r.E_n.M_{x.no}$ etc.)	$V = E_r.E_c.E_n(x_{hij}^2) - M_x^2$ $= E_r.E_c.E_n(x_{hij}^2) - M_x^2$	$\dots$

We may write the identities last derived without ambiguity as

$$V(M_{x.r}) + M(V_{x.r}) = V_x \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad (\text{xix})$$

$$V(M_{x.c}) + M(V_{x.c}) = V_x \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad (\text{xx})$$

$$V(M_{x.n}) + M(V_{x.n}) = V_x \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad (\text{xxi})$$





$$V_{zn} = \frac{206}{81} - 2 - \frac{14}{81} - \frac{14}{81} = \frac{16}{81}.$$



To compute  $V(M_{x.n})$  we must assume some definite order of the items in each cell of (i) corresponding to the 3rd criterion of classification, and assume that we have therein arranged them accordingly. We then have the following lay-out :

(v)

Layer	2.2.2			2.4.3			3.1.2			$T_h$	$T_h^2$
	2.3.1			5.1.3			3.6.3				
	4.0.5			6.2.1			5.5.2				
	8	..	..	13	..	..	11	..	..		
1	8	..	..	13	..	..	11	..	..	32	1024
2	..	5	..	..	7	..	..	12	..	24	576
3	..	..	8	..	..	7	..	..	7	22	484
Totals										78	2084

From (v) we derive

$$S_n = \frac{2084}{9}; S_n - S = \frac{56}{9};$$

$$V(M_{x.n}) = \frac{56}{243}.$$

To derive  $M(V_{x.nc})$ ,  $M(V_{x.nr})$  and hence  $V_{zr}$  or  $V_{zc}$ , we shall need to rearrange the items of (i) alternatively as below.

(vi)

	$i = 1$	$i = 2$	$i = 3$
$h = 1$	2.2.4	2.5.6	3.3.5
$h = 2$	2.3.0	4.1.2	1.6.5
$h = 3$	2.1.5	3.3.1	2.3.2

(vii)

	$h = 1$	$h = 2$	$h = 3$
$j = 1$	2.2.3	2.4.1	2.3.2
$j = 2$	2.5.3	3.1.6	1.3.3
$j = 3$	4.6.5	0.2.5	5.1.2

Square Cell Totals ( $T_{hi}^2$ ).			Total.
64	169	121	354
25	49	144	218
64	49	49	162
Total			734

Square Cell Totals ( $T_{hj}^2$ ).			Total.
49	49	49	147
100	100	49	249
225	49	64	338
Total			734

Whence we obtain

$$S_{nc} = \frac{1}{3}(734); S_{nr} = \frac{1}{3}(734);$$

$$27 \cdot M(V_{x.nc}) = 294 - \frac{1}{3}(734); 27 \cdot M(V_{x.nr}) = 294 - \frac{1}{3}(734),$$

$$\therefore M(V_{x.nc}) = \frac{148}{81}; M(V_{x.nr}) = \frac{148}{81};$$

$$V_{zr} = \frac{76}{9} \text{ and } V_{zc} = \frac{76}{9}.$$

## EXERCISE 11.05

1. Check the dual relation defined by  $V(M_r) + M(V_r) = V = V(M_c) + M(V_c)$  with respect to the following sets of scores first by direct computation and then by the sum of squares schema, i.e.

$$S_q - S = cr \cdot V; S_c - S = cr \cdot V(M_c); S_r - S = cr \cdot V(M_r);$$

$$cr \cdot M(V_c) = S_q - S_c \quad \text{and} \quad cr \cdot M(V_r) = S_q - S_r.$$

(a)	(b)	(c)
1.3.5	2.1.4.0	2.1.9.5.6
4.9.2	3.5.2.1	4.3.3.2.2
6.8.7	2.2.7.4	7.2.1.8.8
	3.1.6.3	

2. For each of the above determine

$$V_z = V - V(M_c) - V(M_r) = M(V_c) + M(V_r) - V.$$

3. By recourse to the  $E$  notation of this section show that

$$E(x_{rc} - M_r - M_c + M)^2 = V_z.$$

Check this result by direct computation w.r.t. the foregoing numerical examples.

*Hint.*—Put  $(x_{rc} - M_r - M_c + M)^2 = [(x_{rc} - M_r) - (M_c - M)]^2$ .

4. Determine the parameters  $V_x$ ,  $V(M_{x.r})$ ,  $V(M_{x.c})$ ,  $V(M_{x.n})$ ,  $M(V_{x.rc})$ ,  $M(V_{x.nc})$  and  $M(V_{x.nr})$  for the following set-up by direct computation in accordance with definition and by recourse to the sum of squares schema:

$$S_q - S = ncr \cdot V; S_c - S = ncr \cdot V(M_c); S_r - S = ncr \cdot V(M_r); S_n - S = ncr \cdot V(M_n)$$

$$S_q - S_{cr} = ncr \cdot M(V_{cr}); S_q - S_{nc} = ncr \cdot M(V_{nc}); S_q - S_{nr} = ncr \cdot M(V_{nr})$$

2.2	4.1	6.3	1.1
1.3	5.4	7.5	0.2
4.2	2.5	4.5	2.3

5. By recourse to the  $E$  notation show that

$$E(M_{rc} - M_r - M_c + M)^2 = V_{zn}.$$

*Hint.*—Remember that the parameter of any particular dimension of a grid is a constant w.r.t. an  $E$  operator in any other dimension, so that

$$E_r(V_{x.c}) = V_{x.c} = E_n(V_{x.c}); E_c(V_{x.r}) = V_{x.r} = E_n(V_{x.r});$$

$$E_r(V_{x.n}) = V_{x.n} = E_c(V_{x.n}); E_n(V_{x.rc}) = V_{x.rc}.$$

6. Use the data of Example 4 to determine  $V_{zc}$  and  $V_{zr}$  defined by analogy with  $V_{zn}$ , i.e.

$$V_{zc} = M(V_r) + M(V_n) - M(V_{nr}) - V;$$

$$V_{zr} = M(V_c) + M(V_n) - M(V_{nc}) - V.$$

Show that the results are numerically consistent with the alternative definitions:

$$V_{zc} = E(M_{nr} - M_r - M_n + M)^2;$$

$$V_{zr} = E(M_{nc} - M_c - M_n + M)^2.$$



## 11.06 ADDITION OF COVARIANCE

It is possible to extend the use of the symbolism of 11.02 to more than 2 dimensions. Thus our concern may be with a function  $F_{abc}$  of 3 score-sets  $x_a, x_b, x_c$ , in which case we may write

$E_{c.ab}(F_{abc})$  = mean value of  $F_{abc}$  for all values of  $c$  when both  $a$  and  $b$  remain constant.

$E_{b.a}.E_{c.ab}(F_{abc})$  = mean value of  $F_{abc}$  for all values of  $b$  and  $c$  when  $a$  remains constant.

$E_a.E_{b.a}.E_{c.ab}(F_{abc})$  = mean value of  $F_{abc}$  for all values of  $a, b$  and  $c$ .

In this symbolism the operation of extracting the grand mean is

$$E_a.E_{b.a}.E_{c.ab}(\dots) \equiv E(\dots) \equiv E_a.E_{c.a}.E_{b.ac}(\dots) \quad . \quad . \quad (i)$$

$$E_b.E_{a.b}.E_{c.ab}(\dots) \equiv E(\dots) \equiv E_b.E_{c.b}.E_{a.bc}(\dots) \quad . \quad . \quad (ii)$$

$$E_c.E_{a.c}.E_{b.ac}(\dots) \equiv E(\dots) \equiv E_c.E_{b.c}.E_{a.bc}(\dots) \quad . \quad . \quad (iii)$$

We may also need to combine the symbolic conventions of 11.02 and 11.05 to cover a case of special interest, e.g. when we have  $c$  sets of paired scores  $x_a$  and  $x_b$ , for each set of which we can assign a covariance. We may then denote :

- (i) by  $E_{ab.c}$  the operation of extracting the mean of a function of both sets of border-scores in one and the same set ;
- (ii) by  $E_c$  the operation of extracting the mean of any parameter (e.g.  $V_{a.c}$  or  $V_{b.c}$ ) of a set.

We should then write

$$E \equiv E_c.E_{ab.c}.$$

In this case, the criteria of classification in the A- and B-dimensions of the 3-dimensional grid are quantitative, always being defined by the border-scores ; but the criterion of classification in the C-dimension is qualitative. For a case of special interest, we may write the covariance of the border-scores in a single set as

$$E_{ab.c}(x_a - M_{a.c})(x_b - M_{b.c}) = Cov(x_{a.c}, x_{b.c}) = E_{ab.c}(x_{a.c}.x_{b.c}) - M_{a.c}.M_{b.c} \quad . \quad (iv)$$

Its mean value for all sets will then be

$$E_c.Cov(x_{a.c}, x_{b.c}) = M.Cov(x_a, x_b) = E(x_a.x_b) - E_c(M_{a.c}.M_{b.c}) \quad . \quad (v)$$

For the covariance of the paired values of  $x_a, x_b$  treated as a whole, we must write

$$E(x_a - M_a)(x_b - M_b) = Cov(x_a, x_b) = E(x_a.x_b) - M_a.M_b \quad . \quad (vi)$$

Now there are  $c$  pairs of mean values  $M_{a.c}, M_{b.c}$  from which we may form

$$E_c(M_{a.c} - M_a)(M_{b.c} - M_b) = Cov(M_{a.c}, M_{b.c}) = E_c(M_{a.c}.M_{b.c}) - M_a.M_b \quad (vii)$$

From (v)-(vii) we thus obtain a tautology of a 3-dimensional set-up reminiscent of (xxvi) in 11.02, viz. :

$$Cov(x_a, x_b) = Cov(M_{a.c}, M_{b.c}) + M.Cov(x_a, x_b) \quad . \quad . \quad (viii)$$

We can express this in a form involving regression coefficients in virtue of (viii) and (x) in 11.03, viz. for regression of the B-score on the A-score :

$$k_{ba} = \frac{Cov(x_a, x_b)}{V_a} ; k_{ma} = \frac{Cov(M_{a.c}, M_{b.c})}{V(M_{a.c})} ; k_{ba.c} = \frac{Cov(x_{a.c}, x_{b.c})}{V_{a.c}} \\ \therefore k_{ba}.V_a = k_{ma}.V(M_{a.c}) + M(k_{ba.c}.V_{a.c}) \quad . \quad . \quad (ix)$$

If each of the  $c$  sets of paired values  $x_a, x_b$  is divisible into  $r$  sub-sets, we may proceed to obtain expressions comparable to (viii) by recourse to the symbolic convention :

$$E_c \cdot E_r \cdot E_{ab \cdot cr}(\dots) \equiv E(\dots) \equiv E_r \cdot E_c \cdot E_{ab \cdot rc}(\dots) \quad (x)$$

Equation (ix) holds good only if regression is linear ; but (viii) is true of any set of numbers which we can pair off in layers of a 3-dimensional grid. Whether the layers are of equal size, i.e. whether the number ( $r_i$ ) of pairs in the  $i$ th layer is the same as in any other, is immaterial. For purposes of computation, we may suitably lay out our figures in 2 dimensions as below. The reader may test the relation involved by recourse to other collections of numbers arranged likewise.

Block.	$U$		$V$		$W$	
	$x_{a \cdot u}$	$x_{b \cdot u}$	$x_{a \cdot v}$	$x_{b \cdot v}$	$x_{a \cdot w}$	$x_{b \cdot w}$
	3	5	10	12	1	3
	2	2	8	9	3	6
	4	7	5	6	2	5
	2	4	3	3	1	2
Totals	11	18	26	30	7	16
Means	2.75	4.5	6.5	7.5	1.75	4.0

The total of the 12  $A$ -scores is  $(11 + 26 + 7) = 44$  ; and that of the  $B$ -scores is  $(18 + 30 + 16) = 64$ . For the grand means we have therefore

$$M_a = \frac{11}{3} \quad \text{and} \quad M_b = \frac{16}{3}.$$

The lay-out for the computation of Covariances is as follows :

	$x_{a \cdot u} \cdot x_{b \cdot u}$	$x_{a \cdot v} \cdot x_{b \cdot v}$	$x_{a \cdot w} \cdot x_{b \cdot w}$	$M_{a \cdot u} \cdot M_{b \cdot u}$
	15	120	3	$198 \div 16$
	4	72	18	$780 \div 16$
	28	30	10	$112 \div 16$
	8	9	2	....
Totals . . .	55	231	33	$545 \div 8$
Means . . .	13.75	57.75	8.25	$545 \div 24$

The mean value of all cross products is

$$\frac{55 + 231 + 33}{12} = \frac{319}{12}.$$

Whence the total Covariance is

$$\frac{319}{12} - \frac{11}{3} \cdot \frac{16}{3} = \frac{506}{72} \quad (xi)$$

The covariance of the means is

$$\frac{545}{24} - \frac{11}{3} \cdot \frac{16}{3} = \frac{227}{72} \quad (xii)$$





## 11.07 SUMMATION BY FIGURATE SERIES

In Vol. I (Chapter 1) we have seen that it is possible to obtain expressions for certain power series relevant to determination of zero moments by recourse to figurate number series. To derive results we shall later use in determining moments of a distribution, in particular summation of products of reduced factorials, it is also convenient to make use of the properties of the family of figurate numbers ( $s = 3$  in Fig. 86) to which the unit, natural numbers, triangular numbers and tetrahedral numbers belong. For the foregoing we may use the symbols  ${}^0F_n$ ,

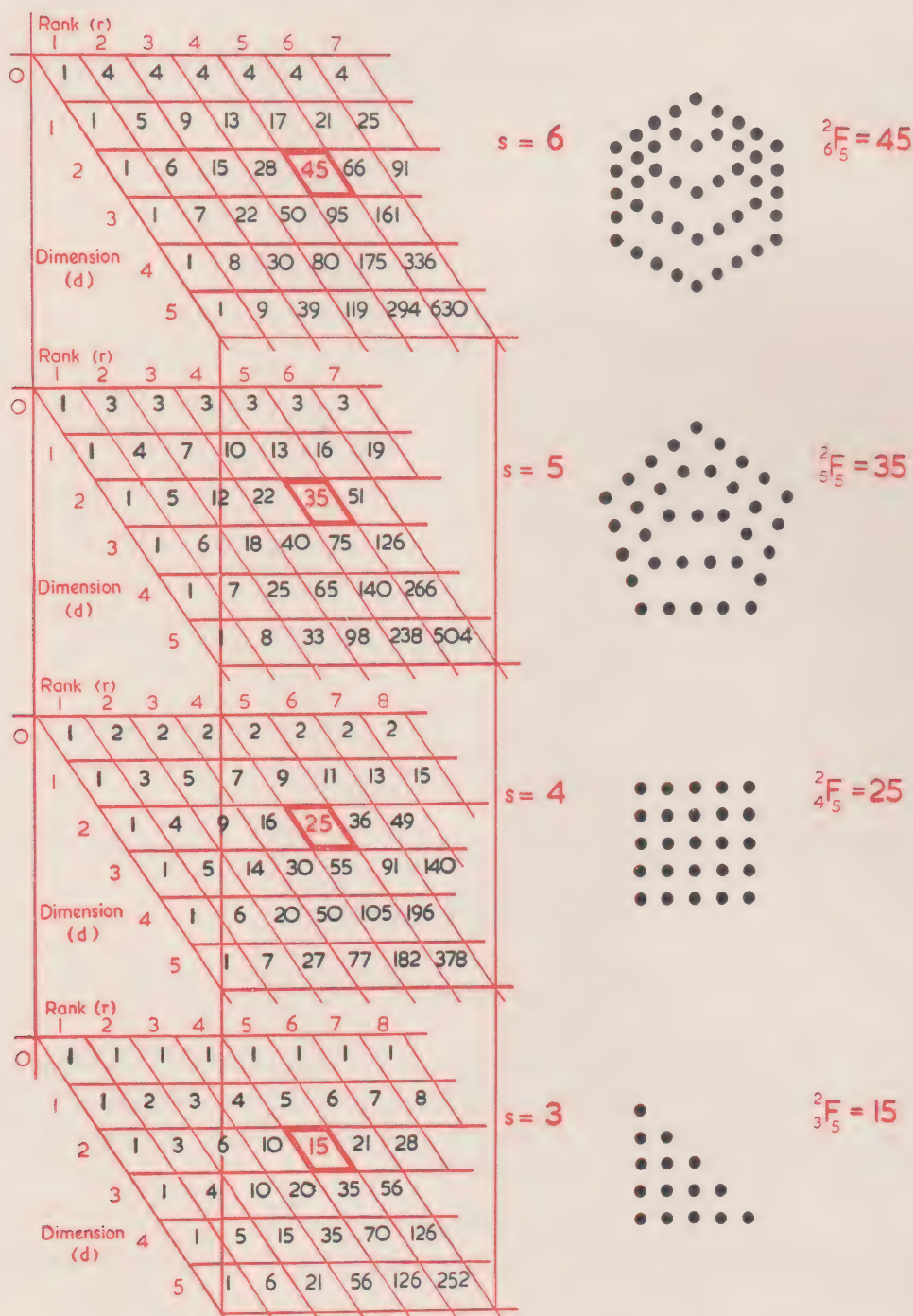


FIG. 86. Figurate Numbers in 3 dimensions.



${}^1F_n, {}^2F_n, {}^3F_n$  severally to denote the term of rank  $n$  and in general  ${}^dF_n$  for that of the  $d$ -dimensional class generated from its predecessor of  $(d - 1)$  dimensions in accordance with the same additive law, so that for positive integral values of  $d$  and  $n$ :

$$\sum_{r=1}^{r=n} {}^d F_r = {}^{d+1} F_n = \sum_{r=0}^{r=n} {}^d F_r \quad . \quad . \quad . \quad . \quad . \quad (\text{i})$$

The general expression for the family is a reduced factorial, *viz.* :

$${}^dF_n = \frac{(n+d-1)^{(d)}}{d!} = (n+d-1)_{(d)} \quad . \quad . \quad . \quad . \quad (\text{ii})$$

If we extend these series into the domain of negative integers the law of summation is (ix) of 1.02, viz. :

$$\sum_{r=-m}^{r=n} {}^d F_r = {}^{d+1} F_n + (-1)^d \cdot {}^{d+1} F_{m-d+1} \quad . \quad . \quad . \quad (\text{iii})$$

Hence if  $k > 2$  and  $n \geq k$

$$\sum_{r=n-k+1}^{r=n-k+1} {}^k F_r = {}^{k+1} F_{n-k+1} + (-1)^k \cdot {}^{k+1} F_{-1}.$$

Since  ${}^{k+1}F_{-1} = 0$  for all positive values of  $k$

$$\sum_{r=-k+1}^{n-k+1} {}^k F_r = {}^{k+1} F_{n-k+1} = \frac{(n+1)^{(k+1)}}{(k+1)!} \quad . \quad . \quad . \quad . \quad (\text{iv})$$

Now we may write for a sum of factorial powers of the integers

$$\sum_{x=1}^{x=n} x^{(k)} = k! \sum_{x=1}^{x=n} \frac{x^{(k)}}{k!} = k! \sum_{x=0}^{x=n} {}^kF_{x-k+1} = k! \sum_{r=-k+2}^{r=n-k+1} {}^kF_r.$$

Whence from (iv) when  $k \geq 2$

$$\sum_{x=1}^{x=n} x^{(k)} = \frac{(n+1)^{(k+1)}}{(k+1)!} \cdot \cdot \cdot \cdot \cdot \quad (\text{v})$$

When  $k = 1$  or  $2$ , this is evidently true, since

$$\sum_{x=1}^{x=n} x^{(1)} = \sum_{x=1}^{x=n} {}^1F_x = {}^2F_n = \frac{(n+1)^{(2)}}{2};$$

$$\sum_{x=1}^{x=n} x^{(2)} = 2 \sum_{x=1}^{x=n} {}^2F_{x-1} = 2 \sum_{r=0}^{r=n-1} {}^2F_r = 2 \cdot {}^3F_{n-1} = \frac{(n+1)^{(3)}}{3}.$$

Hence (v) is valid for all positive integral values of  $k \leq n$ .

In the positive domain the following relationship is also of fundamental importance :

$${}^d F_r = {}^{r-1} F_{d+1} \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad (vi)$$

In virtue of these identities we may now establish the following theorems relating to the sum of products of reduced factorials :

[illegible]

$$\sum_{x=0}^{x=r} (k+x-1)_{(x)}(m+r-x-1)_{(r-x)} = (k+m+r-1)_{(r)} . \quad . \quad . \quad \text{(viii)}$$





We now put  $z = (c - x)$

$$\sum_{c=0}^{c=n} c_{(x)}(n - c)_{(s-x)} = \sum_{z=0}^{z=(n-s)} {}^x F_{z+1} \cdot {}^{s-x} F_{n-s-z+1}.$$

In virtue of (vi) and of (x) above, the expression on the right is equivalent to

$$\sum_{z=0}^{z=(n-s)} {}^z F_{x+1} \cdot {}^{n-s-z} F_{s-x+1} = {}^{n-s} F_{s+2} = {}^{s+1} F_{n-s+1}.$$

In the notation of reduced factorials

$${}^{s+1} F_{n-s+1} = \frac{(n+1)^{(s+1)}}{(n+1)!}.$$

Hence in accordance with (vii)

$$\sum_{c=0}^{c=n} c_{(x)}(n - c)_{(s-x)} = \frac{(n+1)^{(s+1)}}{(s+1)!} = (n+1)_{(s+1)} \quad \text{. . . . . (xii)}$$

We shall later make use of an extension of (xii) which we can derive by recourse to the identity

$$(c+1)_{(x+1)} = \frac{(c+1)c_{(x)}}{(x+1)},$$

$$\therefore (x+1)(c+1)_{(x+1)} = c \cdot c_{(x)} + c_{(x)} \quad \text{. . . . . (xiii)}$$

In virtue of (xiii) we may thus write

$$\sum_{c=0}^{c=n} c \cdot c_{(x)}(n - c)_{(s-x)} = (x+1) \sum_{c=0}^{c=n} (c+1)_{(x+1)}(n - c)_{(s-x)} - \sum_{c=0}^{c=n} c_{(x)}(n - c)_{(s-x)} \quad \text{(xiv)}$$

To reduce the first term on the right to the same form as (xii), we put  $y = (x+1)$ ,  $z = (s+1)$ ,  $m = (n+1)$  and  $u = (c+1)$  so that  $u = 1$  when  $c = 0$  and  $u = m$  when  $c = n$ , whence

$$\sum_{c=0}^{c=n} (c+1)_{(x+1)}(n - c)_{(s-x)} = \sum_{u=1}^{u=m} u_{(y)}(m - u)_{(z-y)}.$$

Since  $u_{(y)} = 0$  when  $u = 0$

$$\sum_{c=0}^{c=n} (c+1)_{(x+1)}(n - c)_{(s-x)} = \sum_{u=0}^{u=m} u_{(y)}(m - u)_{(z-y)} = (m+1)_{(z+1)},$$

$$\therefore \sum_{c=0}^{c=n} (c+1)_{(x+1)}(n - c)_{(s-x)} = (n+2)_{(s+2)}.$$

Hence in virtue of (xii) above, (xiv) becomes

$$\sum_{c=0}^{c=n} c \cdot c_{(x)}(n - c)_{(s-x)} = (x+1)(n+2)_{(s+2)} - (n+1)_{(s+1)} \quad \text{. . . . . (xv)}$$

## 11.08 THE GENERATING FUNCTION AS A GRID OPERATION

To the beginner the generating functions touched on in Chapter 6, Vol. I somewhat savour of being wise after the event. They cease to have an air of mystery when we recognise them as devices for summarising the operations of the *independence* grid (Fig. 87). When we lay out a

## PACKING UP THE CHESSBOARD

						Raw Score					
						0	1	2	3	4	
						5-fold Sample	$a_0$	$a_1$	$a_2$	$a_3$	$a_4$
						4-fold Sample	$b_0$	$b_1$	$b_2$	$b_3$	....
Sums						Frequencies					
0 1 2 3 4						$a_0 a_1 a_2 a_3 a_4$					
0	0	1	2	3	4	$b_0$	$a_0 b_0$	$a_1 b_0$	$a_2 b_0$	$a_3 b_0$	$a_4 b_0$
1	1	2	3	4	5	$b_1$	$a_0 b_1$	$a_1 b_1$	$a_2 b_1$	$a_3 b_1$	$a_4 b_1$
2	2	3	4	5	6	$b_2$	$a_0 b_2$	$a_1 b_2$	$a_2 b_2$	$a_3 b_2$	$a_4 b_2$
3	3	4	5	6	7	$b_3$	$a_0 b_3$	$a_1 b_3$	$a_2 b_3$	$a_3 b_3$	$a_4 b_3$
						Differences					
						0 1 2 3 4					
0	0	1	2	3	4	0	0	1	2	3	4
1	1	2	3	4	5	1	-1	0	1	2	3
2	2	3	4	5	6	2	-2	-1	0	1	2
3	3	4	5	6	7	3	-3	-2	-1	0	1
						Difference as a Sum					
						0 1 2 3 4					
0	0	1	2	3	4	0	0	1	2	3	4
1	1	2	3	4	5	-1	-1	0	1	2	3
2	2	3	4	5	6	-2	-2	-1	0	1	2
3	3	4	5	6	7	-3	-3	-2	-1	0	1

PGF. of (5+4)-fold Sample						PGF. of Raw Score Difference					
Row Score-Sum						w.r.t. 5-fold and 4-fold Samples					
$a_0 t^0 a_1 t^1 a_2 t^2 a_3 t^3 a_4 t^4$						$a_0 t^0 a_1 t^1 a_2 t^2 a_3 t^3 a_4 t^4$					
$b_0 t^0$	$a_0 b_0 t^0$	$a_1 b_0 t^1$	$a_2 b_0 t^2$	$a_3 b_0 t^3$	$a_4 b_0 t^4$	$b_0 t^0$	$a_0 b_0 t^0$	$a_1 b_0 t^1$	$a_2 b_0 t^2$	$a_3 b_0 t^3$	$a_4 b_0 t^4$
$b_1 t^1$	$a_0 b_1 t^1$	$a_1 b_1 t^2$	$a_2 b_1 t^3$	$a_3 b_1 t^4$	$a_4 b_1 t^5$	$b_1 t^1$	$a_0 b_1 t^1$	$a_1 b_1 t^2$	$a_2 b_1 t^3$	$a_3 b_1 t^4$	$a_4 b_1 t^5$
$b_2 t^2$	$a_0 b_2 t^2$	$a_1 b_2 t^3$	$a_2 b_2 t^4$	$a_3 b_2 t^5$	$a_4 b_2 t^6$	$b_2 t^2$	$a_0 b_2 t^2$	$a_1 b_2 t^3$	$a_2 b_2 t^4$	$a_3 b_2 t^5$	$a_4 b_2 t^6$
$b_3 t^3$	$a_0 b_3 t^3$	$a_1 b_3 t^4$	$a_2 b_3 t^5$	$a_3 b_3 t^6$	$a_4 b_3 t^7$	$b_3 t^3$	$a_0 b_3 t^3$	$a_1 b_3 t^4$	$a_2 b_3 t^5$	$a_3 b_3 t^6$	$a_4 b_3 t^7$

FIG. 87. Probability Generating Function for Score-Sum and Score-Difference as a convention for labelling the cells of the chessboard.

grid with border scores 0, 1, 2 . . . etc. with corresponding frequencies  $u_0, u_1, u_2$  . . . etc. and  $v_0, v_1, v_2$  . . . etc., a cell whose score-sum entry is  $s = a + b$  is one whose frequency entry is  $y_{x(s-x)} = u_x \cdot v_{s-x}$ ; and the total frequency of  $s$  for the 2-fold sample is the right-left descending diagonal sum of all such cell entries, as set forth in the following schema for the score-sum ( $s$ ) of unit samples from each of two 4-class universes with score range 0-3:

		0	1	2	3
		$u_0$	$u_1$	$u_2$	$u_3$
0	$v_0$	$s = 0$ $y_{00} = u_0 v_0$	$s = 1$ $y_{10} = u_1 v_0$	$s = 2$ $y_{20} = u_2 v_0$	$s = 3$ $y_{30} = u_3 v_0$
1	$v_1$	$s = 1$ $y_{01} = u_0 v_1$	$s = 2$ $y_{11} = u_1 v_1$	$s = 3$ $y_{21} = u_2 v_1$	$s = 4$ $y_{31} = u_3 v_1$
2	$v_2$	$s = 2$ $y_{02} = u_0 v_2$	$s = 3$ $y_{12} = u_1 v_2$	$s = 4$ $y_{22} = u_2 v_2$	$s = 5$ $y_{32} = u_3 v_2$
3	$v_3$	$s = 3$ $y_{03} = u_0 v_3$	$s = 4$ $y_{13} = u_1 v_3$	$s = 5$ $y_{23} = u_2 v_3$	$s = 6$ $y_{33} = u_3 v_3$

As they appear in the grid lay-out the frequency cell entries  $y_{x(s-x)} (= u_x \cdot v_{s-x})$  lie diagonally as on the left below. We can bring all terms of a diagonal referable to the same score-sum  $s$  into line vertically by sliding the rows as on the right:



$y_{00}$	$y_{10}$	$y_{20}$	$y_{30}$	$y_{00}$	$y_{10}$	$y_{20}$	$y_{30}$	$\dots$	$\dots$	$\dots$
$y_{01}$	$y_{11}$	$y_{21}$	$y_{31}$	$\dots$	$y_{01}$	$y_{11}$	$y_{21}$	$y_{31}$	$\dots$	$\dots$
$y_{02}$	$y_{12}$	$y_{22}$	$y_{32}$	$\dots$	$\dots$	$y_{02}$	$y_{12}$	$y_{22}$	$y_{32}$	$\dots$
$y_{03}$	$y_{13}$	$y_{23}$	$y_{33}$	$\dots$	$\dots$	$\dots$	$y_{03}$	$y_{13}$	$y_{23}$	$y_{33}$
<i>Totals</i>				$Y_0$	$Y_1$	$Y_2$	$Y_3$	$Y_4$	$Y_5$	$Y_6$

The formula for the frequency of the score-sum  $s$  is then

$$Y_s = \sum_{x=0}^{x=s} y_{x(s-x)}.$$

This exhibits the result of applying the product rule as a lay-out on all fours with the familiar algorithm of multiplication. Indeed, the earliest commercial arithmetics set out gridwise the procedure for multiplication with the Hindu-Arabic numerals; and indicated the diagonal totals at the margins. Each such diagonal sum is then a factor of the corresponding power of 10; and we can make the procedure more explicit by attaching to each border-score frequency  $y_x$  a dummy factor  $t^x$  whose exponent is the corresponding score  $x$ . When we then apply the product rule, the cell frequency  $y_s$  carries along with it a factor  $t^s$  whose exponent is the corresponding score-sum cell entry  $s$ . The chessboard operation for the score-sum of any 2 independent samples then assumes the compact form:

$$\begin{array}{c}
 u_0 t^0 \quad u_1 t^1 \quad u_2 t^2 \quad u_3 t^3 \\
 \begin{array}{c} v_0 t^0 \\ v_1 t^1 \\ v_2 t^2 \\ v_3 t^3 \end{array}
 \begin{array}{|c|c|c|c|}
 \hline
 u_0 v_0 t^0 & u_1 v_0 t^1 & u_2 v_0 t^2 & u_3 v_0 t^3 \\
 \hline
 u_0 v_1 t^1 & u_1 v_1 t^2 & u_2 v_1 t^3 & u_3 v_1 t^4 \\
 \hline
 u_0 v_2 t^2 & u_1 v_2 t^3 & u_2 v_2 t^4 & u_3 v_2 t^5 \\
 \hline
 u_0 v_3 t^3 & u_1 v_3 t^4 & u_2 v_3 t^5 & u_3 v_3 t^6 \\
 \hline
 \end{array}
 \equiv
 \begin{array}{|c|c|c|c|}
 \hline
 y_{00} t^0 & y_{10} t^1 & y_{20} t^2 & y_{30} t^3 \\
 \hline
 y_{01} t^1 & y_{11} t^2 & y_{21} t^3 & y_{31} t^4 \\
 \hline
 y_{02} t^2 & y_{12} t^3 & y_{22} t^4 & y_{32} t^5 \\
 \hline
 y_{03} t^3 & y_{13} t^4 & y_{23} t^5 & y_{33} t^6 \\
 \hline
 \end{array}
 \end{array}$$

If the border frequencies in this lay-out refer to a unit sample distribution from different universes of 4 score classes, we define as a *probability generating function* of one or other universe each with the range 0-3:

$$(u_0 t^0 + u_1 t^1 + u_2 t^2 + u_3 t^3) \quad \text{and} \quad (v_0 t^0 + v_1 t^1 + v_2 t^2 + v_3 t^3).$$

We then specify a corresponding probability generating function of the 2-fold sample score-sum as:

$$\begin{aligned}
 & (Y_0 t^0 + Y_1 t^1 + Y_2 t^2 + Y_3 t^3 + Y_4 t^4 + Y_5 t^5 + Y_6 t^6) \\
 & = (u_0 t^0 + u_1 t^1 + u_2 t^2 + u_3 t^3)(v_0 t^0 + v_1 t^1 + v_2 t^2 + v_3 t^3).
 \end{aligned}$$

If we take the samples from the same indefinitely large 4-class universe without replacement or from any 4-class universe with replacement, the above becomes

$$Y_0 t^0 + Y_1 t^1 + Y_2 t^2 + Y_3 t^3 + Y_4 t^4 + Y_5 t^5 + Y_6 t^6 = (u_0 t^0 + u_1 t^1 + u_2 t^2 + u_3 t^3)^2.$$

By successive application of the chessboard operation, the appropriate *p.g.f.* of the  $r$ -fold sample score sum is

$$(u_0 t^0 + u_1 t^1 + u_2 t^2 + u_3 t^3)^r = Y_0 t^0 + Y_1 t^1 + Y_2 t^2 + \dots + Y_{3r} t^{3r}.$$

In any of the foregoing expressions the coefficient ( $u_x$ ,  $v_x$ ,  $Y_x$ ) of  $t^x$  is the frequency of the score  $x$ . By inserting the dummy factor  $t^x$  in the operation for deriving the score-sum distribution, we can pick out the frequency of a particular score-sum without invoking the chessboard lay-out.





If  $a_x$  is the frequency of the score  $x_a$  in the  $a$ -fold sample distribution, it is also that of the mean (or proportionate) score ( $x_a \div a$ ), and we can adapt the foregoing definition of the *p.g.f.* of the score-sum to specify the mean (or proportionate) score frequency by substituting the dummy factor  $t^{\frac{x}{a}}$  for  $t^x$ , so that  $a_x$  is then the coefficient of the  $t$ -term whose exponent is the mean score itself, ie.

$$G(x_a) = \sum_{x=0}^{x=\infty} a_x t^{\frac{x}{a}} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (iv)$$

We should then write for the multiple toss of the cubical die

$$G_2(s_p) = \frac{1}{6^2}(t^{\frac{1}{2}} + 2t + 3t^{\frac{3}{2}} + \dots)$$

We have so far assumed that our unit sample scores increase from zero by unit steps. If they increase from  $m$  by equal steps  $\Delta m$ , we may interpret  $u_x$  as the frequency of the score  $m + x\Delta m$  in the *p.g.f.*

$$G_n = u_0 t^m + u_1 t^{m+\Delta m} + u_2 t^{m+2\Delta m} \dots u_n t^{m+n\Delta m} \quad (v)$$

In the  $a$ -fold score-sum distribution the minimum score will be  $am$  and the corresponding  $p.g.f.$  will be

$$G(x_a) = a_0 t^{am} + a_1 t^{am+\Delta m} + a_2 t^{am+2\Delta m} \dots \text{etc.} \quad . \quad . \quad . \quad (\text{vi})$$

For the corresponding proportionate or mean score, the *p.g.f.* will then be

$$G\left(\frac{x_a}{a}\right) = a_0 t^m + a_1 t^{m+\frac{\Delta m}{a}} + a_2 t^{m+\frac{2\Delta m}{a}} \dots \text{etc.} \quad . \quad . \quad . \quad (\text{vii})$$

*Example 4.*—The scores on the faces of a tetrahedral die are 2, 5, 5, 8, i.e.  $2 + 0(3)$ ,  $2 + 1(3)$  and  $2 + 2(3)$  in the ratio 1 : 2 : 1, so that

$$G_u = \frac{1}{4}(t^2 + 2t^5 + t^8) = \frac{t^2}{4}(t^0 + 2t^3 + t^6) = \frac{t^2}{4}(t^0 + t^3)^2.$$

For the  $a$ -fold sample score-sum :

$$G(x_a) = G_u^a = \frac{t^{2a}}{4^a} (t^0 + t^3)^{2a} = (\frac{1}{2}t)^{2a} \sum_{x=0}^{x=2a} (2a)_{(x)} t^{3x}.$$

If the score-sum is 15 in a 3-fold toss,  $a = 3$  and the relevant term in the expansion is the one whose  $t$  exponent is  $(2a + 3x) = 15$ , so that  $x = 3$ , whence

$$a_x = (2a)_{(x)} \cdot 2^{-2a} = \frac{6!}{3!3!} \frac{1}{64} = \frac{5}{16}.$$

\* \* \* \* \*

We may now generalise an important result already obtained in Chapter 7 of Vol. I. We may define a binomial variate in the domain of *representative* scoring as such if the frequencies of score values  $m + x\Delta m$  increasing by equal steps  $\Delta m$  in the range  $m$  to  $(m + k\Delta m)$  tally with terms of the expansion  $(q + p)^k$ .

$$G_n = t^m(q + pt^{\Delta m})^k,$$

$$\therefore G(x_a) = t^{am}(q + pt^{\Delta m})^{ak} \quad \text{and} \quad G(x_a + x_b) = t^{(a+b)m}(q + pt^{\Delta m})^{(a+b)k}.$$

Thus the frequencies of the score-sum  $am + x\Delta m$  in the  $a$ -fold sample tally with successive terms of the expansion  $(q + p)^{ak}$ . We may express this by saying that a binomial variate defines the distribution of the score-sum (and mean score) of samples of any size from a universe of which the u.s.d. is a binomial variate. If the variance of the u.s.d. is  $kpq$ , that of the  $a$ -fold sample score-sum distribution is  $akpq$ . Hitherto, it has been our custom to lay out the grid for the raw-score difference ( $d$ ) of independent  $a$ -fold and  $b$ -fold samples, as below ( $a = 4, b = 3$ ):

	0 $u_0$	1 $u_1$	2 $u_2$	3 $u_3$	4 $u_4$
0 $v_0$	$d = 0$ $y_{00} = u_0v_0$	$d = 1$ $y_{10} = u_1v_0$	$d = 2$ $y_{20} = u_2v_0$	$d = 3$ $y_{30} = u_3v_0$	$d = 4$ $y_{40} = u_4v_0$
1 $v_1$	$d = -1$ $y_{01} = u_0v_1$	$d = 0$ $y_{11} = u_1v_1$	$d = 1$ $y_{21} = u_2v_1$	$d = 2$ $y_{31} = u_3v_1$	$d = 3$ $y_{41} = u_4v_1$
2 $v_2$	$d = -2$ $y_{02} = u_0v_2$	$d = -1$ $y_{12} = u_1v_2$	$d = 0$ $y_{22} = u_2v_2$	$d = 1$ $y_{32} = u_3v_2$	$d = 2$ $y_{42} = u_4v_2$
3 $v_3$	$d = -3$ $y_{03} = u_0v_3$	$d = -2$ $y_{13} = u_1v_3$	$d = -1$ $y_{23} = u_2v_3$	$d = 0$ $y_{33} = u_3v_3$	$d = 1$ $y_{43} = u_4v_3$

The rule for the individual cell entries  $y_{ij} = u_i v_j$  is as before but the rule of diagonal summation for the particular difference  $d$  is different from the rule for the sum ( $s$ ) being

$$\sum_{j=0}^{j=d} u_{d+j} v_j = Y_d = \sum_{j=0}^{j=d} y_{(d+j)j}.$$

If the maximum value of  $u_i$  is  $u$  and that of  $v_j$  is  $v$  the range of  $d$  is from  $-v$  up to  $+u$ , as from  $-3$  to  $+4$  in the foregoing schema. For which we may write the generating function of the difference as

$$Y_{-3}t^{-3} + Y_{-2}t^{-2} + Y_{-1}t^{-1} + Y_0t^0 + Y_1t^1 + Y_2t^2 + Y_3t^3 + Y_4t^4.$$

More generally, for the difference  $d = (x_a - x_b)$  of two independent variates

$$G(x_a - x_b) = \sum_{d=-v}^{d=u} Y_d t^d.$$

Now we may write the difference  $d = (x_a - x_b)$  as a sum, viz.:  $d = x_a + (-x_b)$ . If, as Fig. 3, we lay out our row border-scores as negative values reversing the sign of the exponent of the attached dummy  $t$  accordingly, we may define a new g.f.:

$$G(-x_b) = b_0t^0 + b_1t^{-1} + b_2t^{-2} + b_3t^{-3} \dots \text{etc.}$$

As before we write the g.f. of the column border-scores

$$G(x_a) = a_0t^0 + a_1t^1 + a_2t^2 + a_3t^3 \dots \text{etc.}$$

Whence we have

$$\begin{aligned} G(x_a) \cdot G(-x_b) &= \sum_{d=-v}^{d=u} \sum_{j=0}^{j=d} a_{d+j} \cdot b_j \cdot t^d \\ &= \sum_{d=v}^{d=u} Y_d \cdot t^d, \end{aligned}$$

$$\therefore G(x_a) \cdot G(-x_b) = G(x_a - x_b) \quad . \quad . \quad . \quad (\text{viii})$$



For the unit sample difference universe of  $(n + 1)$  score classes we may write for brevity

$$G_u = \sum_{x=0}^{x=\infty} u_x t^x \quad \text{and} \quad G_{-u} = \sum_{x=0}^{x=\infty} u_x t^{-x}.$$

The distribution of the  $b$ -fold sample is obtained by expanding  $G_u^b$  of which the terms are identical with those of  $G_{-u}^b$  if we reverse the sign of the exponent of the dummy  $t$ . If  $x_b$  is the  $b$ -fold sample score from such a universe we may therefore write

$$G(-x_b) = G_{-u}^b.$$

As before,  $G_u^a$  is the g.f. of the distribution of the  $a$ -fold sample score ( $x_a$ ), so that  $G_u^a = G(x_a)$  and

$$G(x_a)G(-x_b) = G(x_a - x_b) = G_u^a \cdot G_{-u}^b.$$

If the samples come from different universes we may write this in a more general form as

$$\begin{aligned} G(x_a - x_b) &= G_u^a \cdot G_{-v}^b \\ &= (u_0 + u_1 t^1 + u_2 t^2 \dots)^a (v_0 + v_1 t^{-1} + v_2 t^{-2} \dots)^b \end{aligned} \quad \text{(ix)}$$

*Example 5.*—For the distribution of the raw-score difference between  $a$ - and  $b$ -fold samples from an indefinitely large 2-fold universe

$$\begin{aligned} G(x_a) &= (q + pt)^a; \quad G(-x_b) = (q + pt^{-1})^b; \\ G(x_a - x_b) &= (q + pt)^a (q + pt^{-1})^b. \end{aligned}$$

If  $a = 3$ ,  $b = 2$ , we may set forth the operation as below :

$$\begin{array}{r} q^3 t^0 + 3q^2 p t^1 + 3q p^2 t^2 + p^3 t^3 \\ q^2 t^0 + 2q p t^{-1} + p^2 t^{-2} \\ \hline q^5 t^0 + 3q^4 p t^1 + 3q^3 p^2 t^2 + q^2 p^3 t^3 \\ 2q^4 p t^{-1} + 6q^3 p^2 t^0 + 6q^2 p^3 t^1 + 2q p^4 t^2 \\ q^3 p^2 t^{-2} + 3q^2 p^3 t^{-1} + 3q p^4 t^0 + p^5 t^1 \\ \hline \end{array}$$

Whence we derive the following distribution (left) of the score difference in agreement with the chess-board lay-out (right) below :

out (right) below.

		0	1	2	3		
		$q^3$	$3q^2p$	$3qp^2$	$p^3$		
+ 3	$q^2p^3$	0	$q^2$	0	1	2	3
+ 2	$3q^3p^2 + 2qp^4$			$q^5$	$3q^4p$	$3q^3p^2$	$q^2p^3$
+ 1	$3q^4p + 6q^3p^3 + p^5$			- 1	0	1	2
0	$q^5 + 6q^3p^2 + 3qp^4$			$2q^4p$	$6q^3p^2$	$6q^2p^3$	$2qp^4$
- 1	$2q^4p + 3q^2p^3$			- 2	- 1	0	1
- 2	$q^3p^2$			$q^3p^2$	$3q^2p^3$	$3qp^4$	$p^5$
		1	$2qp$				
		2	$p^2$				

The results summarised in (viii)–(ix) refer to a raw-score difference. They are easily adaptable to the description of the distribution of a score deviation or to a proportionate score, since the only function of the exponent of  $t$  is to label the score itself. To make them do the task required, all we therefore have to do is to label  $t^x$  so that  $x$  is in fact the score which is our concern. Thus we substitute  $t^{x-M}$  for  $t^x$  in the expressions involved in (viii) if our concern is with the score deviation, and  $t^{x/a}$  for  $t^x$  and  $t^{-x/b}$  for  $t^{-x}$  if our concern is with the proportionate score.

For the proportionate score difference  $\left(\frac{x_a}{a} - \frac{x_b}{b}\right)$  of  $a$ -fold and  $b$ -fold samples from an infinite 2-class universe, we therefore have

$$G\left(\frac{x_a}{a} - \frac{x_b}{b}\right) = (q + pt^{\frac{1}{a}})^a (q + pt^{-\frac{1}{b}})^b \quad \dots \quad (x)$$

For the unit sample from an infinite 2-class universe  $M = p$ ,  $X = -p$  when  $x = 0$  and  $X = (1 - p) = q$  when  $x = 1$ . We therefore write the g.f. of the u.s.d. as

$$G_u = (qt^{-p} + pt^q) = t^{-p}(q + pt),$$

$$\therefore G_r(s) = t^{-rp}(q + pt)^r.$$

*Example 6.*—For the heart-score deviation of the 3-fold sample with replacement from a full pack ( $p = \frac{1}{4}$ ):

$$G_u = t^{-\frac{3}{4}}\left(\frac{3}{4} + \frac{1}{4}t\right) = \frac{(3 + t)}{4t^{\frac{3}{4}}},$$

$$G_u^3 = \frac{(3 + t)^3}{4^3 t^{\frac{9}{4}}} = \frac{1}{64}(27t^{-\frac{9}{4}} + 27t^{\frac{1}{4}} + 9t^{\frac{5}{4}} + t^{\frac{9}{4}}).$$

Thus the distribution is

$X$	$-\frac{3}{4}$	$\frac{1}{4}$	$\frac{5}{4}$	$\frac{9}{4}$
$Y$	$\frac{27}{64}$	$\frac{27}{64}$	$\frac{9}{64}$	$\frac{1}{64}$

The rule for adapting the form of the *p.g.f.* to take account of *Change of Scale and Origin* is simple. Consider the following score distributions which differ w.r.t. scale and origin alone:

Score A	.	.	$m$	$m + a$	$m + 2a$	$m + 3a$	$m + 4a$	$m + 5a$	...
Score B	.	.	$q$	$q + b$	$q + 2b$	$q + 3b$	$q + 4b$	$q + 5b$	...
Frequency	.	.	$u_0$	$u_1$	$u_2$	$u_3$	$u_4$	$u_5$	

We may write

$$G_u(A) = \sum_0^\infty u_x \cdot t^{m+ax} = t^m \sum_0^\infty u_x \cdot t^{ax} = t^{m-k} \sum_0^\infty u_x \cdot t^{k+ax};$$

$$G_u(B) = \sum_0^\infty u_x \cdot t^{q+bx} = t^q \sum_0^\infty u_x \cdot t^{bx} = t^{q-k} \sum_0^\infty u_x \cdot t^{k+bx}.$$

Thus the effect of multiplying the  $G_u$  by  $t^{m-k}$  or  $t^{q-k}$  is to change the origin from  $m$  to  $(m + k)$  or from  $q$  to  $(q + k)$  as the case may be. We may reduce both expressions to a form involving the *p.g.f.* of the distribution with unit scale and zero origin by putting

$$t^a = h, \quad t = h^{\frac{1}{a}} \quad \text{and} \quad t^b = g, \quad t = g^{\frac{1}{b}}.$$

We then have

$$G_u(A) = h^{\frac{m}{a}} \sum_0^\infty u_x \cdot h^x \quad \text{and} \quad G_u(B) = g^{\frac{q}{b}} \sum_0^\infty u_x \cdot g^x.$$

### Symmetrical Distributions

It is a property of symmetrical distributions that the distribution of the raw-score sum of  $a$ -fold and  $b$ -fold independent samples has the same form as that of the raw-score of the difference from the same universe, i.e. the only difference between the two being referable to the origin.



Consider the following symmetrical u.s.d. of a universe of 7 score classes :

Score	.	$m$	$m + q$	$m + 2q$	$m + 3q$	$m + 4q$	$m + 5q$	$m + 6q$
Frequency	.	$u_0$	$u_1$	$u_2$	$u_3$	$u_4 = u_2$	$u_5 = u_1$	$u_6 = u_0$

The *p.g.f.* of the u.s.d. is

$$\begin{aligned} G_u &= \sum_0^6 u_x t^{m+qx} = t^m \sum_0^6 u_x t^{qx} \\ &= t^m (u_0 t^0 + u_1 t^q + u_2 t^{2q} + u_3 t^{3q} + u_2 t^{4q} + u_1 t^{5q} + u_0 t^{6q}) \\ &= t^{m+3q} (u_0 t^{3q} + u_1 t^{2q} + u_2 t^q + u_3 t^0 + u_2 t^{-q} + u_1 t^{-2q} + u_0 t^{-3q}). \end{aligned}$$

For the distribution of negative scores we may write

$$\begin{aligned} G_{-u} &= \sum_0^6 u_x t^{-(m+qx)} = t^{-m} \sum_0^6 u_x t^{-qx} \\ &= t^{-m} (u_0 t^0 + u_1 t^{-q} + u_2 t^{-2q} + u_3 t^{-3q} + u_2 t^{-4q} + u_1 t^{-5q} + u_0 t^{-6q}) \\ &= t^{-(m+3q)} (u_0 t^{3q} + u_1 t^{2q} + u_2 t^q + u_0 t^0 + u_2 t^{-q} + u_1 t^{-2q} + u_0 t^{-3q}). \end{aligned}$$

In accordance with the product rule the *p.g.f.* of the raw-score sum  $s = (x_a + x_b)$  of independent  $a$ -fold and  $b$ -fold samples is

$$G(s) = G_u^a \cdot G_u^b = t^{(a+b)(m+3q)} (u_0 t^{3q} + u_1 t^{2q} \dots \text{etc.})^{a+b}.$$

That of the raw-score difference  $d = (x_a - x_b)$  is

$$G(d) = G_u^a \cdot G_{-u}^b = t^{(a-b)(m+3q)} (u_0 t^{3q} + u_1 t^{2q} \dots \text{etc.})^{a+b}.$$

Thus we have

$$G(s) = t^{2b(m+3q)} G(d).$$

As we have seen, the only effect of the left-hand factor in the expression on the right is to change the origin of the distribution. This result is easy to confirm by recourse to the chessboard device. If the distribution is symmetrical, diagonal summation from left to right downwards is equivalent to diagonal summation downwards from right to left in the *square* grid ; and the student should be able to interpret the change of origin in terms of the distribution of the *score deviations* by drawing it.

### EXERCISE 11.08

1. A card pack contains only equal numbers of cards of the following denominations : ace of hearts, 2 of clubs, 3 of spades, the player's score being the total number of pips regardless of suit. Write down the distribution of the 3-fold sample score-sum and that of the difference between 2-fold samples on the assumption of *replacement*, and check the result by recourse to the chessboard procedure.

2. Derive by means of the *p.g.f.* the frequency of the following mean scores for a 3-fold toss of the tetrahedral dice with faces as specified :

Mean Score	Faces of die
4	3, 4, 5, 5
5	1, 5, 5, 9
10	2, 6, 10, 14

Check the results by the grid procedure.

3. Specify distributions of the difference between both the total score and the mean score of the 3-fold and the 2-fold toss of each of the dice in Example 2 above.

4. By use of the g.f. involving the dummy factor  $t^x$  establish the following conclusions w.r.t. sampling with replacement from a 2-class universe (e.g. red or black cards of a full pack) when  $p = \frac{1}{2} = q$ :

- (a) the raw-score difference distribution about the mean for  $a$ -fold and  $b$ -fold samples is the same as that of the  $(a + b)$ -fold score-sum;
- (b) the proportionate score difference is identical with that of the raw-score about its mean.

5. When  $p$  and  $q$  are not equal in Example 3, show that

- (a) the raw-score deviation difference distribution is identical with that of the proportionate score if the size of the sample is equal;
- (b) the distribution about the mean of the score-sum for samples of  $2a$  cards is the same as the distribution of the sum of the differences between  $a$  pairs.

6. For an infinite 3-class universe of score values  $-1, 0$  and  $1$ , write out the distribution of the 3-fold sample mean score on the assumption that the ratios of the score class frequencies are (a)  $1 : 1 : 1$ ; (b)  $1 : 2 : 1$ ; (c)  $1 : 4 : 1$ . Check by the chessboard procedure.



## CHAPTER 12

# MODELS OF BIVARIATE UNIVERSES

### 12.00 MEANING OF THE MODELS

ABOUT the turn of the century Karl Pearson adapted methods of line fitting prescribed by Legendre and Gauss for the evaluation of physical constants based on laboratory experiments to the description of concomitant measurements of relatives. The ostensible end in view was to develop certain confused and erroneous beliefs about inheritance propounded by Francis Galton, in deference to whose *mystique* the Gaussian *Method of Least Squares* asserted its claims in a new context under a new name as *regression*. Pearson's new contribution was the announcement of a measure of association commonly called the correlation coefficient or more precisely the product-moment index. With its aid he claimed to have established on a firm footing Galton's so-called *Law of Ancestral Inheritance*. This generalisation is meaningful in one sense which is true but too trite to have any claims to novelty or to utility. In any other sense, it is demonstrably false.

Through Bowley the new evangel of correlation spread to the social sciences, expounded against a background of geometrical concepts which defy any attempt to make explicit the manifold circumstances in which co-variation may arise. Not unnaturally the social sciences have therefore laboured under a load of misconceptions from which the revival of Mendel's experimental method rescued the study of heredity in plants and animals. To make explicit circumstances relevant to a correct assessment of observed correlation is therefore a task of no mean importance. It is indeed a simple matter, if we examine different types of model situations. By examining one such class of models in Chapter 9 of Vol. I, we have seen that correlation in the statistical sense entails no necessary conclusions about the causal nexus involved in the events recorded. We shall appreciate this more clearly if we now take stock of some very diverse situations in which correlation can arise. Such is our chief concern in what follows ; but the situations we are about to explore may prove to be misleading if we do not clearly appreciate in what sense each model dealt with is a *universe* in contradistinction to a *sample* such as we meet in sociological or biological research.

In our first approach to statistical theory it is appropriate to regard the structure of the universe (e.g. a *card pack*) as the source of our information about samples (e.g. *hands at bridge*) drawn therefrom ; but we have anticipated a different viewpoint in so far as we have found it (a) necessary to draw a sharp distinction between *die* or *lottery wheel* models and *card pack* or *urn* models in Chapter 2 of Vol. I ; (b) convenient to speak of the score-frequency specification of the universe as the *unit sample distribution*. We shall be better able to appreciate the lesson of the models dealt with in this chapter if we first re-examine our use of the terms universe and sample.

From a *static* viewpoint we may usefully distinguish between universes of 3 kinds :

- (i) *discrete and finite*, if there is a finite number of score classes each with a finite number of identical score values, e.g. a full pack of 52 cards of which 13 (*score value* 1) are hearts and 39 (*score value* 0) are of other suits ;
- (ii) *discrete and infinite*, if we pool an infinite number of full card packs, in which event there is a finite number of score classes each with an infinite number of identical score values, subject to the understanding that the ratio of two such infinite numbers is specifiable and finite ;



- (iii) *continuous*, in the sense that the number of score classes is infinite and the number of items in each class is infinite though not necessarily equivalent on that account.

The last named is a convenient mathematical fiction which at least serves a useful purpose as a means of simplifying laborious computation with sufficient accuracy for practical purposes, e.g. when we invoke the normal curve to specify the distribution of large samples (Chapter 3 in Vol. I) from a 2-class universe of type (ii) above. Likewise it is often a convenient device for specifying the structure of a discrete universe of the same type when the number of score classes is very large. Whether it is more than a fiction is open to philosophic doubt; and we are on solid ground only if we confine our attention in this context to (i) and (ii).

If we define our universe as both finite and discrete, we cannot specify the results of sampling from it unless we agree at the outset concerning whether the sampling process does or does not involve replacement of each item chosen before taking another. If we impose the condition of replacement, the distinction between (i) and (ii) ceases to be relevant from a mathematical viewpoint, since sampling with replacement from one full card pack is equivalent to sampling without replacement from an infinite number of full card packs.

When our model universe is an urn or a card pack, we are always free to regard it as an entity in its own right on the assumption that we are free to sample one way or the other; but we cannot appropriately conceive the model universe of the die or lottery wheel in this way. By its very nature, such a model is a widow's cruse. However often we toss a cubical die, it is still possible to score a six at the next trial. Thus, the structure of the model is such that we must in effect impose the replacement condition on the sampling process and therefore conceive the universe of the model as a universe of type (ii); but in making any such statement about our model universe we have changed our viewpoint. If we view a penny as a *static* entity, we are entitled to regard it as a 2-class universe with 2 score values; but such a picture of the universe might lead us to wrong conclusions about its behaviour, if the penny were biased. To visualise it correctly with that end in view, we must think of it as a universe *in action*. We have then to conceive it as a 2-class universe with an infinite number of score values, the ratio of alternative score values being unity if the penny is unbiased.

To say that we must so conceive it as a universe *in action* is to say that we have reversed the more naive procedure of deducing the nature of the sampling distribution from the structure of the universe. We are now conceiving the nature of the universe in terms of the sampling process. This is the readjustment we have to make, if we seek to visualise sampling in what we shall later define as a *correlation universe*. To do this, let us recall a simple example of the class of model situations dealt with in Chapter 9. The umpire tosses a coin twice, each of two players (*A* and *B*) toss once, each adding the umpire's score (heads) to his (or her) own individual scores. We may summarise grid-wise the players' joint score distribution in an *indefinitely large number* of trials as below (as in 11.01, p. 429):

		$x_a$			
		0	1	2	3
$x_b$	0	1	1	0	0
	1	1	3	2	0
	2	0	2	3	1
	3	0	0	1	1



Each set of border-scores and the corresponding row or column totals (not shown above) of such a grid defines a *univariate* (single score) unit sample distribution, being thus a summary of the relative frequencies of recording each score value in an infinite number of trials. In other words, each border-score distribution defines a particular universe of scores. The cell entries of the grid exhibit how often a score of one set turns up with a score of the other set in the long run. As such, they summarise the relative frequencies of the paired scores  $(x_a, x_b)$  in an infinite number of trials, each cell entry (divided by the grand total) being the probability of getting a particular paired score at a single trial. This is what we mean when we speak of a *bivariate unit-sample* distribution. Since it summarises the outcome of an infinite number of independent trials, such a distribution describes a universe from which we may extract samples of any number on the assumption that the particular paired score-value obtained at one trial (*unit sample* of paired scores) does not affect the paired score-value obtained at the next. To say that the universe is infinite in this context is, of course, consistent with saying that it contains a finite number of classes. In the same sense, we have seen that the universe of the common cubical die is both discrete and infinite. The number of faces defines the relative frequencies of six classes each with an infinite number of items available for withdrawal as sample values.

Only a sample composed of 16 or some exact multiple of 16 paired scores could be specifiable by exactly the same proportionate cell entries as the bivariate universe of the bonus model mentioned above ; and this would be a very rare occurrence. In general, the sample structure prescribed by, say, 8 successive trials may be specified by filling in the cells with integers up to a total of 8 (or exact multiples of one-eighth up to a total of unity) with the proviso that certain cells whose corresponding theoretical frequencies are zero remain empty. In the above, these zero cells are, of course, defined by 0.2, 0.3, 1.3, 2.0, 3.0, 3.1. Thus 3 possible 8-fold samples are as below :

2	0	.	.
0	2	0	.
.	0	2	0
.	.	0	2

$$r_{ab} = +1$$

0	0	.	.
0	2	2	.
.	2	2	0
.	.	0	0

$$r_{ab} = 0$$

0	0	.	.
0	0	4	.
.	4	0	0
.	.	0	0

$$r_{ab} = -1$$

Of many possible 8-fold samples one might choose, the above illustrate the possibility that an actual sample may ring the changes on values of  $r_{ab}$  from perfect negative through zero to perfect positive correlation, and hence bring into focus a twofold problem about a correlation universe :

- (a) what sample parameter is an unbiased estimate of the product-moment index  $r_{ab}$  in the sense that  $(r - 1)s^2 = r\sigma^2$  defines the unbiased  $r$ -fold sample estimate (p. 304, Vol. I) of the variance ( $\sigma$ ) of a univariate unit sample distribution, i.e. universe of single scores ?
- (b) how can we define the sample distribution of  $r_{ab}$  or other characteristic parameter, e.g.  $k_{ba}$  or  $Cov(x_a, x_b)$  ?

At this stage, we shall not attempt to answer these questions, stating them merely to emphasise the importance of the distinction between  $r_{ab}$ , etc. conceived as parameters of a *correlation universe* (unit-sample bivariate distribution) and as parameters of a particular sample of observations. In the exposition of the models in this chapter our concern is with the specification of





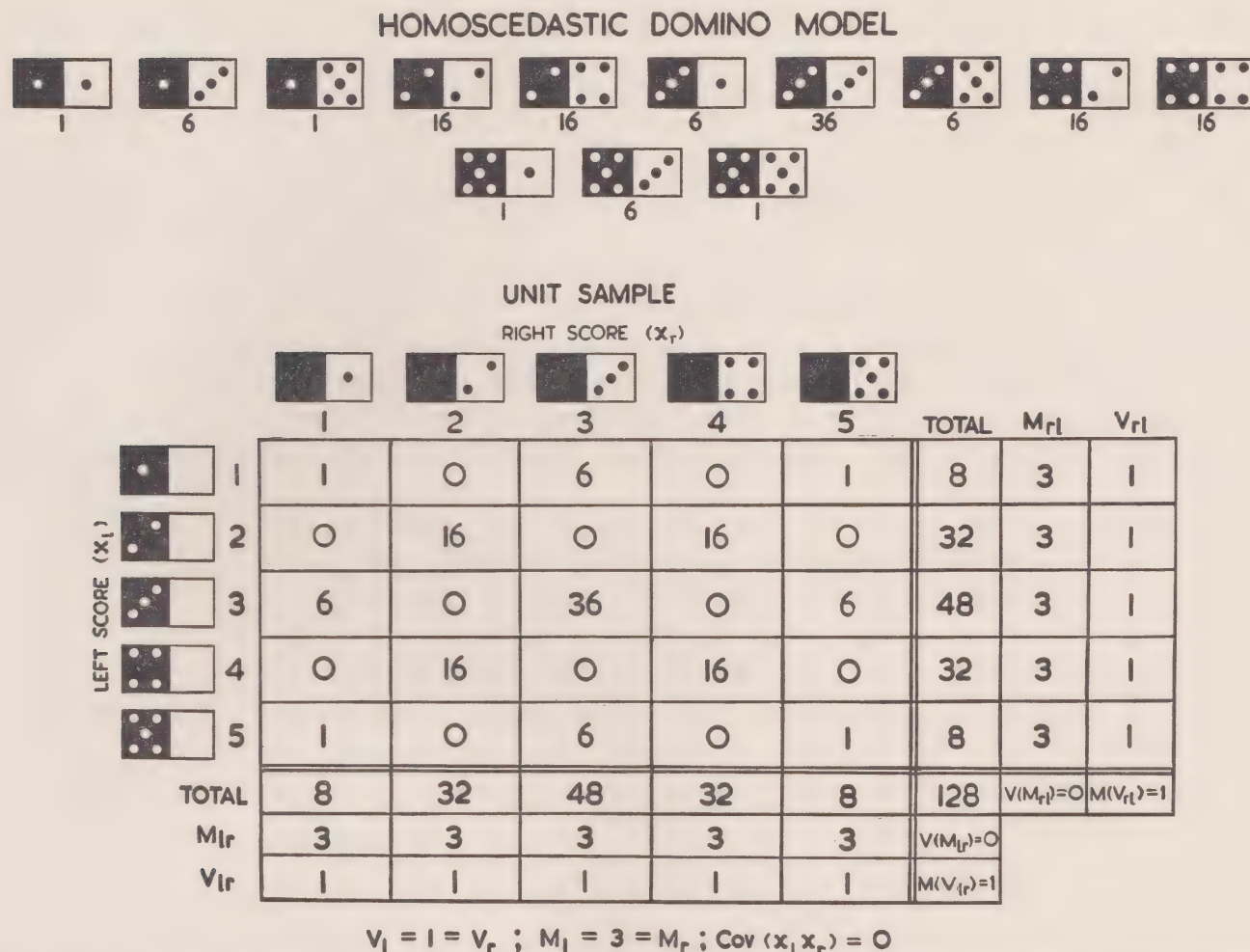


FIG. 89. A bivariate universe which has zero covariance and homoscedasticity (equality of variance within arrays) in both dimensions.

It is important to recognise the implications of the fact that the two distributions set forth above each represent the distributions of *independent* unit samples, i.e. *trials*. When we speak of one as a correlation universe and the other as a universe of independence the distinction refers to the way in which the *A*-scores are distributed w.r.t. the *B*-scores (or *vice versa*) among unit samples of paired scores ; but when we prescribe the relative frequencies of *r*-fold samples from a universe of either sort, we do so on the assumption that *one trial is independent of another in conformity with the usual chessboard procedure*. In short, we can visualise our bivariate universe as both finite and discrete, if we impose the condition of replacement on the process of sampling therefrom.

We may suppose that the umpire has a large box of dominoes of 10 denominations, the two halves of each domino being respectively white with black pips and *vice versa*. Instead of writing down the score at each trial, the umpire may draw from the box a domino with the appropriate number of white pips on black for the total score of player *A* and of black pips on white for the total score of player *B*. As the contest goes on the pile of each of the ten types grows, and as the number of games becomes indefinitely large the proportions of dominoes in each pile become ever closer to the distribution assigned by the grid. From a formal point of view, the long-run result would be exactly the same if we recorded the result of taking dominoes with replacement one at a time from a box of only 10 dominoes of the same ten types in the same proportions (Fig. 88). This is a convenient visualisation inasmuch as it helps us to see one way of constructing

a distribution of which (Fig. 89) we can predicate both zero covariance and homoscedasticity (equal variance within arrays) without imposing the additional restriction that the two score distributions are truly independent in the statistical sense of the term.

One feature of Fig. 88 calls for special comment because it introduces a device which will later (Chapter 18) steer us through a maze of difficulties in the theory of sampling from a bivariate universe. Since each of the dominoes, i.e. players' joint score chips, of Fig. 88 records the result of a single trial, we can visualise the trial as a grid with a single cell entry in the appropriate row and column, otherwise constructed like the summarising grid of the unit sample distribution (u.s.d.). Since each trial is independent of another, we can operate with such unit grids in accordance with the chessboard lay-out, successively deriving the distribution of 2-fold, 3-fold, etc. samples by applying the product rule, at each stage specifying the composition of the  $(r + 1)$ -fold sample of frequency so defined by adding cell by cell the entries of the unit grid to that of the  $r$ -fold sample grid. Thus the following shows the generation of a particular 2-fold sample :

$$\begin{array}{ccc}
 \begin{array}{|c|c|c|c|} \hline 1 & . & . & . \\ \hline . & . & . & . \\ \hline . & . & . & . \\ \hline . & . & . & . \\ \hline \end{array} & + & \begin{array}{|c|c|c|c|} \hline . & . & . & . \\ \hline . & 1 & . & . \\ \hline . & . & . & . \\ \hline . & . & . & . \\ \hline \end{array} \\
 f_{00} = \frac{1}{16} & & f_{11} = \frac{3}{16} \\
 \\ 
 \begin{array}{|c|c|c|c|} \hline . & . & . & . \\ \hline . & 1 & . & . \\ \hline . & . & . & . \\ \hline . & . & . & . \\ \hline \end{array} & + & \begin{array}{|c|c|c|c|} \hline 1 & . & . & . \\ \hline . & . & . & . \\ \hline . & . & . & . \\ \hline . & . & . & . \\ \hline \end{array} \\
 f_{11} = \frac{3}{16} & & f_{00} = \frac{1}{16} \\
 \\ 
 & = & \begin{array}{|c|c|c|c|} \hline 1 & . & . & . \\ \hline . & 1 & . & . \\ \hline . & . & . & . \\ \hline . & . & . & . \\ \hline \end{array} \\
 & & f_{00 \cdot 11} = \frac{3}{128}
 \end{array}$$

### EXERCISE 12.00

1. Each of a pack of cards carries 1, 2 or 3 hearts on one face and 1, 2 or 3 spades on the other in the following proportions :

1H	1S	$p$	2H	1S	$s$	3H	1S	$v$
1H	2S	$q$	2H	2S	$t$	3H	2S	$w$
1H	3S	$r$	2H	3S	$u$	3H	3S	$z$

Examine the properties of the *unit* sample distribution for the following values  $p, q$ , etc. with special reference to the row and column means, the row and column variances, the values of  $\eta_{ab}^2$ ,  $\eta_{ba}^2$  and  $r_{ab}$  :

	$p$	$q$	$r$	$s$	$t$	$u$	$v$	$w$	$z$
(a)	1	0	0	2	4	2	0	0	1
(b)	0	1	0	1	2	1	0	1	0
(c)	1	1	0	2	12	2	0	1	1

2. Work out the distribution of 2-fold samples for (a)-(c) above.



## 12.01 THE UMPIRE BONUS MODEL

In Chapter 9 of Vol. I we have examined a type of correlation which arises between the scores ( $x_a$  and  $x_b$ ) of two players  $A$  and  $B$  in a game of chance, if each receives a variable bonus ( $x_u$ ) from a third player called the umpire. In Chapter 18 we shall rely on the same model to clarify the elements of the statistical procedure called *Factor Analysis*. For that reason, we shall now examine it in more general terms. Fig. 90 exhibits a simple variation of the model, *viz.* :

### UMPIRE BONUS MODEL

	3	4	5	6	7	Total	$M_{ob}$	$V_{ob}$
3	1	2	1	0	0	4	$\frac{140}{35}$	$\frac{1}{2}$
4	3	8	7	2	0	20	$\frac{154}{35}$	$\frac{37}{50}$
5	0	6	13	8	1	28	$\frac{180}{35}$	$\frac{61}{98}$
6	0	0	3	6	3	12	$\frac{210}{35}$	$\frac{1}{2}$
Total	4	16	24	16	4	64	5	1
$M_{ba}$	$\frac{15}{4}$	$\frac{17}{4}$	$\frac{19}{4}$	$\frac{21}{4}$	$\frac{23}{4}$	$\frac{19}{4}$	$M_o = 5$	$V_o = 1$
$V_{ba}$	$\frac{3}{16}$	$\frac{7}{16}$	$\frac{25}{48}$	$\frac{7}{16}$	$\frac{3}{16}$	$\frac{11}{16}$	$M_b = \frac{19}{4}$	$V_b = \frac{11}{16}$

$$M(V_{ob}) = \frac{22}{35}; V(M_{ob}) = \frac{13}{35}; M(V_{ba}) = \frac{7}{16}; V(M_{ba}) = \frac{1}{4}$$

$$M(V_{ob}) + V(M_{ob}) = 1; \quad M(V_{ba}) + V(M_{ba}) = \frac{11}{16}$$

$$\underline{\text{Cov}}(AB) = \frac{1}{2} = V_{11}$$

$$r_{ab}^2 = \frac{4}{11}$$

$$\frac{V(M_{ab})}{V_a} = \frac{13}{35} ; \quad \frac{V(M_{ba})}{V_b} = \frac{4}{11}$$

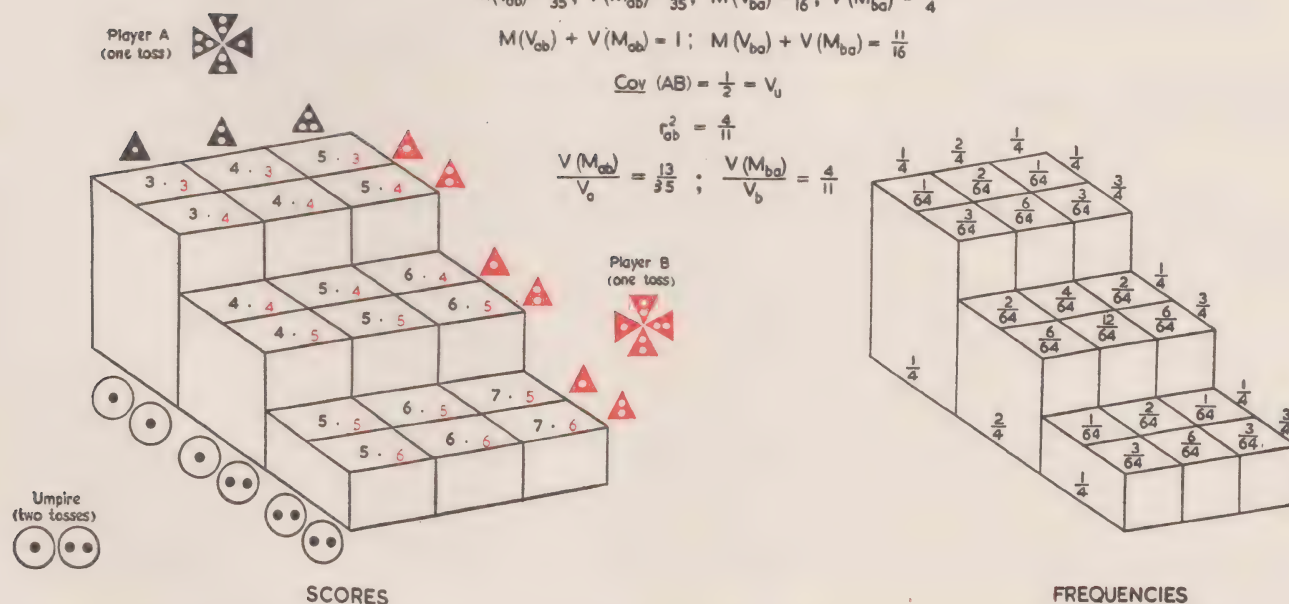


FIG. 90. Umpire Bonus Model—linear regression in one dimension only. The umpire tosses twice the flat, circular die of Fig. 67, Vol. I. Player *A* tosses once the tetrahedral die of Fig. 70. Player *B* tosses once the tetrahedral die of Fig. 73.

- (i) The umpire tosses twice the die of Fig. 67, Chapter 7 ;
- (ii) Player *A* tosses once the die of Fig. 70, Chapter 7 ;
- (iii) Player *B* tosses twice the die of Fig. 73, Chapter 7.

In such a situation we note that the scores ( $x_{a,o}$  and  $x_{b,o}$ ) of the individual players  $A$  and  $B$  before addition of the bonus are strictly independent of one another and of the score of the umpire. We express this by the equations

$$x_a = x_u + x_{a.o} \quad \text{and} \quad x_b = x_u + x_{b.o} \quad . \quad . \quad . \quad . \quad (\text{i})$$





When  $A_u = 1 = B_u$ , this reduces to  $Cov(x_a, x_b) = V_u$  as we have seen in Chapter 9 of Vol. I. For the more general case we may write

$$r_{ab} = \frac{A_u B_u \cdot V_u}{\sqrt{V_a V_b}} = \frac{A_u \sigma_u}{\sigma_a} \cdot \frac{B_u \sigma_u}{\sigma_b} \quad (vii)$$

To determine the covariance of the players' score and that of the umpire we note that

$$\begin{aligned} X_u X_a &= X_u (A_u X_u + A_o X_{a.o}) = A_u X_u^2 + A_o X_u X_{a.o}, \\ \therefore Cov(x_a, x_u) &= A_u E(X_u^2) + A_o E(X_u \cdot X_{a.o}) = A_u \cdot V_u. \end{aligned}$$

Whence we obtain

$$r_{au} = \frac{A_u V_u}{\sigma_a \sigma_u} = A_u \frac{\sigma_u}{\sigma_a} \quad (viii)$$

Similarly, we derive

$$\begin{aligned} r_{bu} &= B_u \frac{\sigma_u}{\sigma_b}, \\ \therefore r_{au} \cdot r_{bu} &= A_u \frac{\sigma_u}{\sigma_a} \cdot B_u \frac{\sigma_u}{\sigma_b} = r_{ab} \quad (ix) \end{aligned}$$

In connexion with the derivation of the equation of partial correlation and the meaning of the "reliability" coefficient in 18.06 below we may usefully elaborate our model, as in the following examples:

*Example 1.*—Umpire  $U$  tosses an ordinary cubical die once. Umpire  $W$  tosses a coin twice (scoring heads as 1 and tails as 0). Player  $A$  tosses the tetrahedral die of Fig. 70 (scores 1, 2, 3 in the ratio 1 : 2 : 1). Player  $B$  tosses once the tetrahedral die of Fig. 73 (scores 1 and 2 in the ratio 1 : 3). The final score of  $A$  is twice his individual score added to 3 times the score of umpire  $U$  and four times the score of umpire  $W$ . The final score of  $B$  is three times his individual score together with twice that of umpire  $U$  and 5 times that of umpire  $W$ , i.e.

$$\begin{aligned} x_a &= 3x_u + 4x_w + 2x_{a.o}; \\ x_b &= 2x_u + 5x_w + 3x_{b.o}. \end{aligned}$$

*Example 2.*—Umpire  $U$  tosses a cubical die once. Player  $A$  first tosses the die of Fig. 67 twice, and then tosses once the die of Fig. 70, counting as his individual score 4 times the result obtained with the circular die added to the result of tossing the tetrahedral die. Player  $B$  tosses the die of Fig. 73 once, and adds to 5 times his score the result of tossing a coin twice, the sum being his individual score. The final score of  $A$  is three times that of the umpire added to his individual score.  $B$  adds to his individual score twice that of the umpire, i.e.

$$\begin{aligned} x_a &= 3x_u + 4x_{as} + x_{ae}; \\ x_b &= 2x_u + 5x_{bs} + x_{be}. \end{aligned}$$

In *Example 1*, we have two independent umpires, and we may write the general pattern as

$$x_a = A_u x_u + A_w x_w + A_o x_{a.o}; \quad x_b = B_u x_u + B_w x_w + B_o x_{b.o} \quad (x)$$

In this set-up

$$\begin{aligned} M_a &= A_u M_u + A_w M_w + A_o M_{a.o}; \quad M_b = B_u M_u + B_w M_w + B_o M_{b.o}; \\ V_a &= A_u^2 V_u + A_w^2 V_w + A_o^2 V_{a.o}; \quad V_b = B_u^2 V_u + B_w^2 V_w + B_o^2 V_{b.o}. \end{aligned}$$





An important generalisation developed in 18.06 below emerges from a modification of this pattern when the player plays against himself by repeating at each trial his prescribed number of tosses with the third die, and adding the result of the single toss of the first and of the second to each. If the player ( $A$ ) does this, he has at each trial 2 scores :

$$x_1 = A_u x_u + A_s x_{a.s} + A_e x_{e.1},$$

$$x_2 = A_u x_u + A_s x_{a.s} + A_e x_{e.2}.$$

This conforms to the pattern of Example 1 with the special condition that  $V_{e.1} = V_e = V_{e.2}$ , and we may write

$$\begin{aligned} \text{Cov}(x_1, x_2) &= A_u^2 V_u + A_s^2 V_s, \\ V_1 &= A_u^2 V_u + A_s^2 V_s + A_e^2 V_e = V_2 \end{aligned} \quad . \quad . \quad . \quad (\text{xvi})$$

Whence we have

$$r_{12} = \frac{A_u^2 V_u + A_s^2 V_s}{A_u^2 V_u + A_s^2 V_s + A_e^2 V_e} \quad . \quad . \quad . \quad (\text{xvii})$$

Our numerical examples in Chapter 9 of Vol. I have shown that linear regression in neither dimension of the grid is a necessary consequence of linear concomitant variation in the domain of the concurrent relationship between the scores of the two players when Example 1 defines the set-up. Contrariwise, they also show that there is always linear regression of the player's score on that of the umpire. To show that this is always true, we shall first consider the case which arises when : (a) the player's individual score and that of the umpire each increase by unit steps from zero upwards ; (b) the definitive equation of the  $A$ -score is  $x_a = x_u + x_{a.o}$ . For this set-up  $\text{Cov}(x_a, x_u) = V_u$  and

$$r_{au} = \frac{V_u}{\sigma_a \sigma_u} = \frac{\sigma_a}{\sigma_u}.$$

If regression is linear, the identity defined by (vii) of 11.04 then implies that

$$k_{au} = r_{au} \frac{\sigma_a}{\sigma_u} = 1 \quad . \quad . \quad . \quad . \quad (\text{xviii})$$

Let us first recall the build-up of the correlation grid for the concomitant score distribution of player and umpire by reference to a simple example, *viz.* : the player  $A$  tosses a penny twice and the umpire four times, the player's total score being the sum of the player's individual score and that of the umpire. If  $p$  is the probability of scoring a success in a single toss, we may denote the distribution as follows :

Score	Frequency	
	Umpire	Player's individual score
0	$u_0 = q^4$	$a_0 = q^2$
1	$u_1 = 4pq^3$	$a_1 = 2pq$
2	$u_2 = 6p^2q^2$	$a_2 = p^2$
3	$u_3 = 4p^3q$	...
4	$u_4 = p^4$	...

For a fixed umpire score of 0, 1, 2, etc. the *total A*-scores run (0, 1, 2), (1, 2, 3), (2, 3, 4), etc., or more generally for a  $U$ -score of  $r$  the range is from  $r$  to  $(r + 2)$  with weighted frequencies  $u_r . a_0$ ,  $u_r . a_1$  and  $u_r . a_2$ . The frequency grid is therefore as overleaf.

		A's Total Score						
		0	1	2	3	4	5	6
Umpire's Bonus	0	$q^6$	$2pq^5$	$p^2q^4$	0	0	0	0
	1	0	$4pq^5$	$8p^2q^4$	$4p^3q^3$	0	0	0
	2	0	0	$6p^2q^4$	$12p^3q^3$	$6p^4q^2$	0	0
	3	0	0	0	$4p^3q^3$	$8p^4q^2$	$4p^5q$	0
	4	0	0	0	0	$p^4q^2$	$2p^5q$	$p^6$
A-score frequency		$q^6$	$6pq^5$	$15p^2q^4$	$20p^3q^3$	$15p^4q^2$	$6p^5q$	$p^6$

When the two sets of independent scores both increase by unit steps from zero origin and the player's total score is simply the sum of his individual score and that of the umpire, we may generalise as below the preceding pattern to accommodate situations involving the use of different dice by the player and the umpire :

		A's Total Score					
		0	1	2	3	4	...
Umpire's Score	0	$u_0a_0$	$u_0a_1$	$u_0a_2$	$u_0a_3$	$u_0a_4$	...
	1	—	$u_1a_0$	$u_1a_1$	$u_1a_2$	$u_1a_3$	...
	2	—	—	$u_2a_0$	$u_2a_1$	$u_2a_2$	...
	3	—	—	—	$u_3a_0$	$u_3a_1$	...

If the player's individual score increases from 0 to  $z$  by unit steps, its mean value is

$$M_{a..o} = a_0(0) + a_1(1) + a_2(2) \dots a_z(z) = \sum_{x=0}^{x=z} a_x \cdot x \quad \dots \quad \text{(xix)}$$

By definition also

$$a_0 + a_1 + a_2 \dots + a_z = \sum_{x=0}^{x=z} a_x = 1.$$

For the  $r$ th row of the grid the mean  $A$ -score ( $M_{a..r}$ ) is

$$\begin{aligned} M_{a..r} &= \frac{u_r \cdot a_0(r) + u_r \cdot a_1(r+1) + u_r \cdot a_2(r+2) \dots u_r \cdot a_z(r+z)}{u_r \cdot a_0 + u_r \cdot a_1 + u_r \cdot a_2 \dots u_r \cdot a_z} \\ &= a_0(r) + a_1(r+1) + a_2(r+2) \dots a_z(r+z) \\ &= r \sum_{x=0}^{x=z} a_x + \sum_{x=0}^{x=z} a_x \cdot x, \\ \therefore M_{a..r} &= r + M_{a..o} \quad \dots \quad \text{(xx)} \end{aligned}$$

In this expression  $r$  is the umpire's score and  $M_{a..o}$  is constant. Thus the mean  $A$ -score of a row is a linear function of the border  $U$ -score. To arrive at this conclusion we have assumed that (a) both the score of the umpire and the individual score of the player increase by unit steps from zero origin ; (b) constants  $A_u$  and  $A_o$  are each equal to unity in the definitive equation

$$x_a = A_u \cdot x_u + A_o \cdot x_{a..o}.$$

We shall remove these restrictions by postulating that

- (a) the constants  $A_u$  and  $A_o$  may have any value ;
- (b) the increments  $\Delta x_u$  and  $\Delta x_{a..o}$  may have any constant value ;
- (c) the origin of the distributions of  $x_u$  and  $x_{a..o}$  may have any value  $m_u$  and  $m_o$  respectively.



$$A_n(m_n + r \cdot \Delta x_n) + A_o(m_o + c \cdot \Delta x_{a..o}).$$
$$K_{c,r} = A_n(m_n + r \cdot \Delta x_n) + A_0 \cdot m_0 \quad . \quad . \quad . \quad . \quad (\text{xxi})$$
[illegible][illegible]
$$M_o = \sum_{c=0}^{c=z} a_c(m_o + c \cdot \Delta x_{a \cdot o}) = m_o + \Delta x_{a \cdot o} \sum_{c=0}^{c=z} a_c \cdot c.$$
$$M_{a,r} = K_{c,r} + A_0 \cdot M_0 - A_0 \cdot m_0.$$
$$M_{a..r} = A_n . m_n + A_0 . M_0 + A_n . \Delta x_n . r \quad . \quad . \quad . \quad (\text{xxiv})$$

From a factual point of view there is a clear-cut distinction between a set-up involving the contribution of 2 umpires ( $U$  and  $W$ ) and one which involves a single umpire ( $Z$ ); but the two situations may be algebraically identical, as is easy to see when: (a) the umpires  $U$  and  $W$  respectively toss the same die  $u$  and  $w$  times; (b) each player adds the actual score of each

umpire to his individual score, so that  $A_u = 1 = B_u$ , etc. and  $A_w = 1 = B_w$ . In these circumstances we may denote the mean and variance of the unit sample distribution of the die of the umpires respectively by  $M_1$  and by  $V_1$ , so that  $M_u = u \cdot M_1$ ;  $V_u = u \cdot V_1$  and  $M_w = w \cdot M_1$ ;  $V_w = w \cdot V_1$ , whence

$$M_u + M_w = (u + w)M_1 \quad \text{and} \quad V_u + V_w = (u + w)V_1.$$

The covariance of the joint distribution of the players is therefore the same as if a single umpire  $Z$  tossed the same die  $z = (u + w)$  times, since then

$$V_z = z \cdot V_1 = \text{Cov}(x_a, x_b) = (u + w)V_1 = V_u + V_w.$$

Let us now remove the restrictions imposed above w.r.t. the nature of the dice and the equal allocation of the bonuses. For simplicity, we may take the origin of all score distributions as the mean, so that replacement implies that

$$A_u X_u + A_w X_w + A_o X_{a.o} = X_a = A_z X_z + A_o X_{a.o} \quad . \quad . \quad (\text{xxv})$$

$$B_u X_u + B_w X_w + B_o X_{b.o} = X_b = B_z X_z + B_o X_{b.o} \quad . \quad . \quad (\text{xxvi})$$

For such a system we may write

$$A_u^2 V_u + A_w^2 V_w + A_o^2 V_{a.o} = V_a = A_z^2 V_z + A_o^2 V_{a.o},$$

$$B_u^2 V_u + B_w^2 V_w + B_o^2 V_{b.o} = V_b = B_z^2 V_z + B_o^2 V_{b.o},$$

$$\therefore A_u^2 V_u + A_w^2 V_w = A_z^2 V_z \quad \text{and} \quad B_u^2 V_u + B_w^2 V_w = B_z^2 V_z \quad . \quad . \quad (\text{xxvii})$$

$$A_u B_u V_u + A_w B_w V_w = \text{Cov}(x_a, x_b) = A_z B_z V_z \quad . \quad . \quad (\text{xxviii})$$

Let us now write

$$A_w^2 V_w = K_a A_u^2 V_u; \quad B_w^2 V_w = K_b B_u^2 V_u; \quad A_w B_w V_w = K_{ab} A_u B_u V_u \quad . \quad . \quad (\text{xxix})$$

From (xxvii) we then have

$$A_u^2(1 + K_a)V_u = A_z^2 V_z \quad \text{and} \quad B_u^2(1 + K_b)V_u = B_z^2 V_z,$$

$$\therefore A_u \sigma_u \sqrt{(1 + K_a)} = A_z \sigma_z \quad \text{and} \quad B_u \sigma_u \sqrt{(1 + K_b)} = B_z \sigma_z \quad . \quad . \quad (\text{xxx})$$

$$\therefore A_z B_z V_z = A_u B_u V_u \sqrt{(1 + K_a)(1 + K_b)}.$$

Also from (xxviii) and (xxix)

$$A_z B_z V_z = A_u B_u V_u (1 + K_{ab}).$$

Hence the condition of replaceability of the two umpires by one umpire without affecting the intercorrelation of the players' score distribution is

$$(1 + K_{ab})^2 = (1 + K_a)(1 + K_b).$$

This must be an identity if  $K_a = K_{ab} = K_b$ , in which case (xxix) implies that

$$\frac{A_w^2 V_w}{A_u^2 V_u} = \frac{B_w^2 V_w}{B_u^2 V_u} = \frac{A_w B_w V_w}{A_u B_u V_u},$$

$$\therefore \frac{A_w}{A_u} = \frac{B_w}{B_u} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (\text{xxxi})$$

Thus we may say that one umpire  $Z$  can replace the two umpires  $U$  and  $W$  if the multiples of the two umpires' scores respectively allocated to player  $A$  and player  $B$  are in the same ratio.



[illegible]

1. As in Example 1 above, umpire  $U$  tosses an ordinary cubical die once. Umpire  $W$  tosses one coin twice (scoring heads as 1 and tails as 0). Player  $A$  tosses the tetrahedral die of Fig. 70 (scores 1, 2, 3 in the ratio 1 : 2 : 1). Player  $B$  tosses once the tetrahedral die of Fig. 73 (scores 1 and 2 in the ratio 1 : 3). The final score of  $A$  is twice his individual score added to 3 times the score of umpire  $U$  and four times the score of umpire  $W$ . The final score of  $B$  is three times his individual score together with twice that of umpire  $U$  and five times that of umpire  $W$ , i.e.

$$Cov(x_a, x_b) = 27.5; r_{ab} = 0.898.$$

$$\text{Cov}(x_a, x_b) = 17.5; r_{ab} = 0.74.$$

- (i) Player  $A$  tosses only the die of Fig. 70 and player  $B$  tosses only the die of Fig. 73.
- (ii) Player  $A$  tosses only the die of Fig. 67 and player  $B$  tosses only the die of Fig. 73.
- (iii) Player  $A$  tosses only the die of Fig. 70 and player  $B$  tosses only the coin.
- (iv) Player  $A$  tosses only the die of Fig. 67 and player  $B$  tosses only the coin.

Compare the results with those of Example 2 with special reference to the variances of the score distributions of the players, and interpret them.

7. An umpire tosses a coin three times scoring heads as successes. Player *A* takes three cards with replacement from a full pack counting the number of hearts as his individual score. Player *B* tosses a cubical die once. Player *A* records as his final score the result of *deducting* three times the score of the umpire from twice his individual score. Player *B* adds to his individual score twice that of the umpire. Investigate the joint distribution of the scores of the two players and that of each player with that of the umpire.

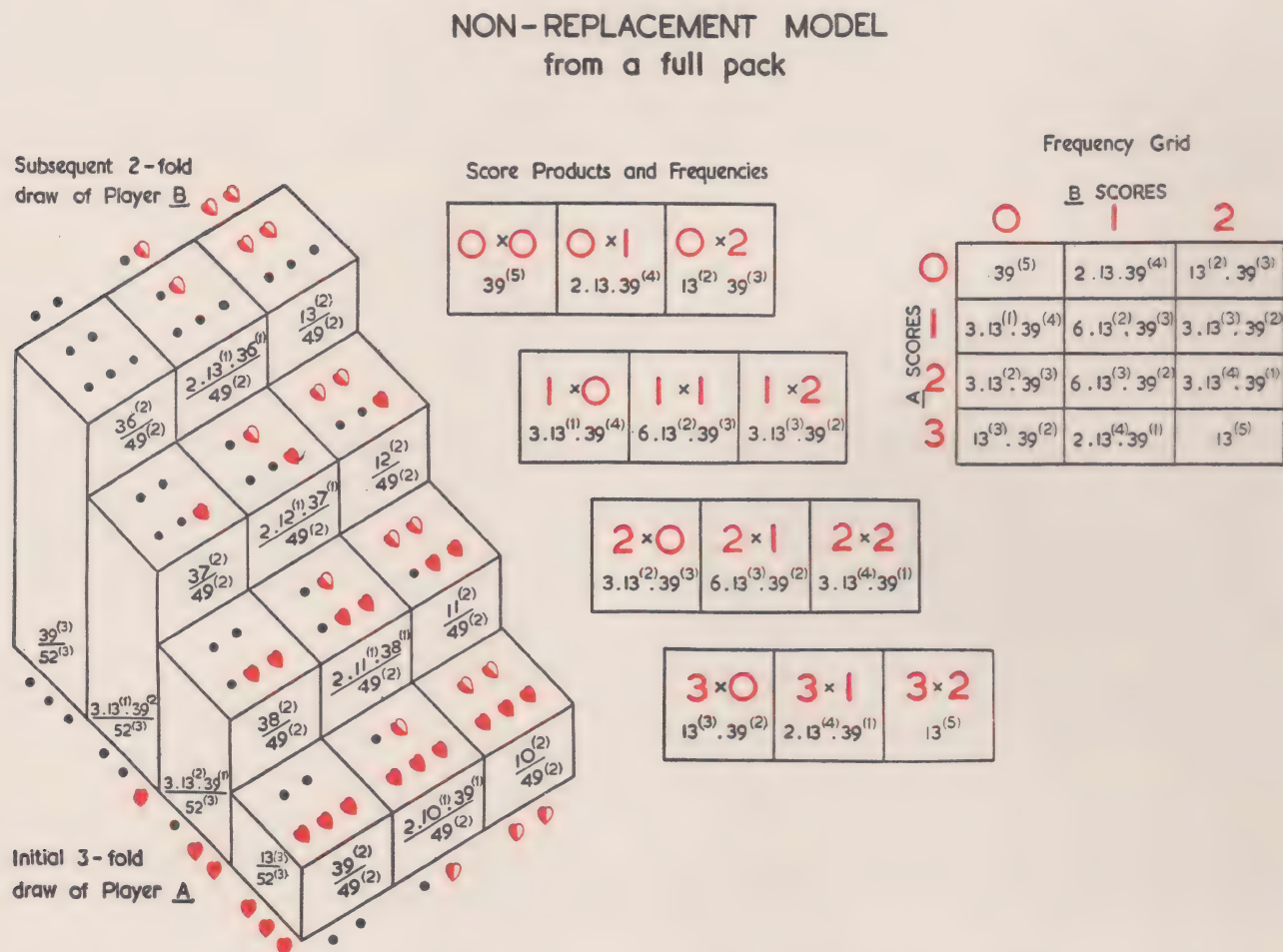


FIG. 91. A Non-replacement Model. Player *A* takes 3 cards from a full pack without replacement and player *B* then takes 2, each recording as his or her score the number of hearts in the sample.

## 12.02 THE NON-REPLACEMENT MODEL

Whether we score by the *taxonomic* or by the *representative* method distinguished in 7.01 of Vol. I, the withdrawal of a sample without replacement limits the possible score value of a sample drawn subsequently from the same universe. If one player takes a sample containing all the highest cards, a second player must evidently have a lower mean score than otherwise. Hence there will be a *negative* correlation between the players' scores. Correlation of this sort is the theme of the models whose properties we shall now explore. A numerical example involving the taxonomic method of scoring will clarify our task:



*Example.*—From a 6-card pack consisting of 2 clubs and 4 hearts, the first player (*A*) simultaneously draws two cards and the second player (*B*) draws three from the residual pack of four. *A*'s heart score (0, 1 or 2) distribution is given by successive terms of  $(2 + 4)^{(2)}/6^{(2)}$ , viz.

<i>A</i> 's score	.	.	.	.	0	1	2
Frequency ( $\times 15$ )	.	.	.	.	1	8	6

The residual packs from which *B* draws are as follows :

<i>A</i> 's heart score	.	.	.	.	0	1	2
Residual :							
hearts	.	.	.	.	4	3	2
clubs	.	.	.	.	0	1	2

If *A*'s heart score is 0, that of *B* is necessarily 3 for the 3-fold draw. If *A*'s heart score is 1, the distribution of *B*'s heart score of 0, 1, 2 or 3 accords with the terms of  $(1 + 3)^{(3)}/4^{(3)}$ . If *A*'s heart score is 2, the appropriate binomial is  $(2 + 2)^{(3)}/4^{(3)}$ . Thus we get a frequency table :

					<i>B's heart score</i>			
					0	1	2	3
when $x_a = 0$	.	.	.	.	0	0	0	1
$x_a = 1$	.	.	.	.	0	0	$\frac{3}{4}$	$\frac{1}{4}$
$x_a = 2$	.	.	.	.	0	$\frac{1}{2}$	$\frac{1}{2}$	0

To obtain the correlation grid we have to weight the above by corresponding frequencies of *A*'s score, and obtain

						<i>B</i>				
						0	1	2	3	Mean
	0	0	0	0	1					3.00
	1	0	0	6	2					2.25
<i>A</i>	2	0	3	3	0					1.50
										$M_a = 1.3$
Mean		0	2	1.3	0.6					$M_b = 2.0$

From the above we obtain

$$\begin{aligned}
 V_a &= 16/45; \quad V_b = 2/5; \\
 k_{ba} &= -3/4; \quad k_{ab} = -2/3; \\
 \text{Cov}(x_a, x_b) &= -4/15 = k_{ba}V_a = k_{ab}V_b; \\
 r_{ab} &= -1/\sqrt{2}.
 \end{aligned}$$

We shall later see that  $r_{ab}$  depends only on the sampling fractions defined by  $n \cdot f_a = a$  and  $n \cdot f_b = b$ , and is in fact  $(f_a \cdot f_b)/(1 - f_a)(1 - f_b) = r_{ab}^2$ . In the above example,

$$\begin{aligned}
 f_a &= 2/6 = \frac{1}{3}; \quad f_b = 3/6 = \frac{1}{2}; \\
 \therefore (f_a \cdot f_b)/(1 - f_a)(1 - f_b) &= \frac{1}{2} = r_{ab}^2.
 \end{aligned}$$

### The Two-class Universe

When we employ the taxonomic method of scoring as in the foregoing example, we may make use of relations we have already obtained in Chapter 3 of Vol. I. For illustrative purposes we assume that *A* and *B* each records the number ( $x_a$  or  $x_b$ ) of hearts in the sample, the proportion

$$\begin{aligned} x_a \cdot M_{ba} &= \frac{b \cdot x_a(np - x_a)}{n - a} = \frac{nb p \cdot x_a}{n - a} - \frac{b \cdot x_a^2}{n - a}, \\ \therefore E_a(x_a \cdot M_{ba}) &= \frac{nb p \cdot M_a}{n - a} - \frac{b}{n - a} E_a(x_a^2) \\ &= \frac{nb p^2}{n - a} - \frac{b}{n - a} (V_a + M_a^2) \\ &= \frac{nb p^2}{n - a} - \frac{a^2 b p^2}{n - a} - \frac{ab}{n - 1} pq \\ &= ab p^2 - \frac{ab}{n - 1} pq. \end{aligned}$$





So we may write

$$\begin{aligned}
 V_{b \cdot a} &= \frac{b(n-a-b)}{(n-a-1)} \cdot \frac{(np-x_a)}{n-a} \cdot \frac{(n-a-np+x_a)}{n-a} \\
 &= \frac{b(n-a-b)}{(n-a)^2(n-a-1)}(np-x_a)(nq-a+x_a), \\
 \therefore V_{b \cdot a} &= \frac{b(n-a-b)}{(n-a)^2(n-a-1)}[n^2pq - anp + (2np-n+a)x_a - x_a^2], \\
 \therefore M(V_{b \cdot a}) &= \frac{b(n-a-b)}{(n-a)^2(n-a-1)}E_a[n^2p - n^2p^2 - anp + (2np-n+a)x_a - x_a^2].
 \end{aligned}$$

In this expression  $E_a(x_a^2) = V_a + M_a^2$ , whence

$$\begin{aligned}
 E_a[n^2p - n^2p^2 - anp + (2np-n+a)x_a - x_a^2] &= pq(n-a)^2 - V_a, \\
 \therefore M(V_{b \cdot a}) &= \frac{b(n-a-b)pq}{n-a-1} - \frac{ab(n-a-b)pq}{(n-a)(n-a-1)(n-1)}.
 \end{aligned}$$

Whence from (ix) and (x)

$$V_b = \frac{b(n-a-b)pq}{n-a-1} - \frac{ab(n-a-b)pq}{(n-a)(n-a-1)(n-1)} + \frac{ab^2pq}{(n-a)(n-1)}.$$

Hence in agreement with (vi) :

$$V_b = \frac{b(n-b)}{n-1}pq.$$

### Representative Scoring

We shall now assume that the pack contains no picture cards. Accordingly we may score the sample by adding the number of pips on each card, e.g. the score of a 3-fold sample consisting of the ace of clubs, the 3 of spades and the 7 of hearts will be 11. Whether the pack contains equal numbers of cards with the same number of pips is immaterial.

As before we assume that player  $A$  draws  $a$  cards from the full pack of  $n$  cards without replacement and player  $B$  draws  $b$  cards without replacement from the residual pack of  $(n-a)$  cards. We shall denote the total score (*score-sum*) of player  $A$  by  $s_a$  and that of player  $B$  by  $s_b$ . We may assign to each individual card a rank referable to the order in which the player draws it if he draws the cards successively, or to their place in a face upwards sequence if drawn simultaneously. We may then label the score of the  $u$ th card drawn as  $x_u$ . The subscript has nothing to do with the numerical value of the score, its function being to label the card as a *removal*, i.e. one which the residual pack does not contain. By definition, then

$$s_a = \sum_{u=1}^{u=a} x_u \quad \text{and} \quad s_b = \sum_{u=a+1}^{u=b} x_u \quad . \quad . \quad . \quad . \quad (xi)$$

If we write the second zero moment of the two score distributions as  $\mu_{2a}$  and  $\mu_{2b}$ , the corresponding variances are

$$V_a = \mu_{2a} - M_a^2 \quad \text{and} \quad V_b = \mu_{2b} - M_b^2 \quad . \quad . \quad . \quad . \quad (xii)$$



In these expressions

$$\begin{aligned} M_a &= E(s_a) = E(x_1 + x_2 + x_3 \dots x_a) \\ &= E(x_1) + E(x_2) \dots + E(x_a) \end{aligned} \quad (xiii)$$

$$\begin{aligned} \mu_{2a} &= E(s_a^2) = E(x_1 + x_2 + x_3 \dots x_a)^2 \\ &= E(x_1^2) + E(x_2^2) + E(x_3^2) \dots \text{etc.} \\ &\quad + 2E(x_1 \cdot x_2) + 2E(x_2 \cdot x_3) \dots \text{etc.} \end{aligned} \quad (xiv)$$

To evaluate  $M_a$  and  $\mu_{2a}$  we must first define  $E(x_u)$  and  $E(x_u^2)$ , i.e. the mean value of the unit score and of its square regardless of the order of choice. With this end in view, we shall denote the score-sum of all the cards in the  $n$ -fold pack by  $S_n$  and the sum of the squares of all the individual card scores as  $S_{2n}$ , so that

$$S_n = \sum_{u=1}^{u=n} x_u \quad \text{and} \quad S_{2n} = \sum_{u=1}^{u=n} x_u^2. \quad (xv)$$

For the corresponding score-sums of all the cards in the residual pack of  $(n - p + 1)$  cards after extraction of a  $(p - 1)$ -fold sample, i.e. immediately before extraction of the  $p$ th card, we may use  $S_{n-p+1}$  and  $S_{2(n-p+1)}$ , so that

$$S_{n-p+1} = \sum_{u=p}^{u=n} x_u \quad \text{and} \quad S_{2(n-p+1)} = \sum_{u=p}^{u=n} x_u^2. \quad (xvi)$$

Let us now consider the result of two consecutive draws, the  $p$ th and the  $(p + 1)$ th. Since the player takes a card from a residual universe of  $(n - p + 1)$  cards at the  $p$ th draw we may write the mean score drawn as

$$E(x_p) = \frac{E(S_{n-p+1})}{n - p + 1}.$$

The operation  $E$  here refers to all possible values  $p$  may take but involves no restriction on the residual universe other than the fact that it contains the card whose score value is  $x_p$ . Hence we may write with equal propriety

$$E(x_p) = E_p(x_p) = \frac{E_p(S_{n-p+1})}{n - p + 1}. \quad (xvii)$$

If we write  $q = (p + 1)$ , the operation  $E_{q \cdot p}$  signifies taking the mean value of the score at the  $q$ th draw after removing the particular card whose score is  $x_p$  at the preceding draw, and this is equivalent to finding the mean score of a unit sample from a residual pack of  $(n - p)$  cards whose score sum is  $S_{n-p+1} - x_p$ . Thus

$$E_{q \cdot p}(x_q) = \frac{S_{n-p+1} - x_p}{n - p}.$$

The mean value of  $(x_q)$  is the weighted mean of the above for all values of  $p$ th draw, i.e.

$$E(x_q) = E_p \cdot E_{q \cdot p}(x_q) = \frac{E_p(S_{n-p+1})}{n - p} - \frac{E_p(x_p)}{n - p}.$$

Whence from (xvii)

$$E(x_a) = \frac{E_p(S_{n-p+1})}{n-p} - \frac{E_p(S_{n-p+1})}{(n-p)(n-p+1)} = \frac{E_p(S_{n-p+1})}{n-p+1},$$

$$\therefore E(x_a) = E(x_p),$$

$$\therefore E(x_{p+1}) = E(x_p).$$

Thus the mean value of the card drawn is independent of choice and therefore  $E(x_u) = E(x_1)$ . If we denote the mean value of the first card drawn from the full pack as  $M_1 = E(x_1)$ , we therefore have

$$M_a = E(x_1) + E(x_2) + \dots + E(x_a) = aM_1. \quad (\text{xviii})$$

Similarly,  $M_b = b \cdot M_1$ . If we substitute  $x_p^2$ ,  $x_q^2$ , and  $S_{2(n-p+1)}$  for  $x_p$ ,  $x_q$  and  $S_{n-p+1}$  in the foregoing argument and write  $E(x_1^2) = \mu_2$ , we derive in the same way

$$E(x_u^2) = \mu_2 = \frac{S_{2n}}{n}. \quad (\text{xix})$$

To complete the evaluation of  $V_a$  and  $V_b$  we also need to determine the mean value of the product terms in (xiv) above. If we write  $x_u \cdot x_w$  as the product of *any* two unit scores of the sample, its mean value is the weighted mean of the product of  $x_u$  and the mean value of  $x_w$  for the same fixed value of  $x_u$ , i.e. in the symbolism of 11.01–11.04:

$$E(x_u \cdot x_w) = E_u[x_u \cdot E_{w \cdot u}(x_w)].$$

To interpret the meaning of the operation  $E_{w \cdot u}$  we recall that the only restriction implicit is that the residual universe does not contain  $x_u$ , and we have already seen that the mean value of  $x_w$  does not depend on order of choice. Hence  $E_{w \cdot u}$  means taking the mean value of  $x_w$  from a pack which does not contain  $x_u$ , i.e. from an  $(n-1)$ -fold residual universe of which the score-sum is  $(S_n - x_u)$ , so that

$$E_{w \cdot u}(x_w) = \frac{S_n - x_u}{n-1}.$$

In this expression  $S_n = nM_1$ , so that

$$E_{w \cdot u}(x_w) = \frac{n}{n-1}M_1 - \frac{x_u}{n-1},$$

$$\therefore x_u \cdot E_{w \cdot u}(x_w) = \frac{n}{n-1}M_1 \cdot x_u - \frac{x_u^2}{n-1},$$

$$\therefore E(x_u \cdot x_w) = \frac{n}{n-1}M_1 E_u(x_u) - \frac{E_u(x_u^2)}{n-1}.$$

Whence from (xix)

$$E(x_u \cdot x_w) = \frac{n}{n-1}M_1^2 - \frac{1}{n-1}\mu_2. \quad (\text{xx})$$

Thus the mean value of the cross products in (xiv) is independent of the order of choice. In (xiv) the number of terms is  $a^2$ , of which  $a$  have the form  $x_u^2$  and  $a(a-1)$  have the form  $x_u \cdot x_w$ , the former being the diagonal terms when we lay out the operation of squaring grid-wise, e.g.:





From (xxii) above we see that

$$V_b = \frac{b(n-b)}{a(n-a)} V_a,$$

$$\therefore r_{ab} = -\frac{\sqrt{ab}}{\sqrt{(n-a)(n-b)}} \quad \text{. . . . . (xxv)}$$

If we denote the sampling fractions by  $f_a$  and  $f_b$  as above (i.e.  $a = nf_a$  and  $b = nf_b$ ),

$$r_{ab} = -\sqrt{\frac{f_a \cdot f_b}{(1-f_a)(1-f_b)}} \quad \text{. . . . . (xxvi)}$$

We may arrive at (xxiii) by an alternative route. If  $a$  is fixed we have

$$M_{b \cdot a} = E_{b \cdot a}(s_b) = E_{b \cdot a}(x_{a+1} + x_{a+2} \dots x_b) = b \cdot E_{b \cdot a}(x_u).$$

In this expression  $E_{b \cdot a}(x_u)$  is the mean value of a  $b$ -fold score-sum from an  $(n-a)$ -fold universe of which the score-sum is  $(S_n - s_a)$ , so that

$$M_{b \cdot a} = \frac{b}{n-a}(S_n - s_a) = \frac{nb}{n-a}M_1 - \frac{b}{n-a}s_a,$$

$$\therefore M_{b \cdot a} - M_b = \frac{nb}{n-a}M_1 - \frac{b}{n-a}s_a - b \cdot M_1,$$

$$\therefore M_{b \cdot a} - M_b = -\frac{b}{n-a}(s_a - M_a) \quad \text{. . . . . (xxvii)}$$

There is therefore linear regression of  $s_b$  on  $s_a$ , and (by the same reasoning) of  $s_a$  on  $s_b$ , and the regression coefficient in (xxvii) is

$$k_{ba} = -\frac{b}{n-a}.$$

In virtue of linear regression, we may write

$$\text{Cov}(s_a, s_b) = k_{ba} \cdot V_a,$$

$$\therefore \text{Cov}(s_a, s_b) = -\frac{b}{n-a} \cdot V_a.$$

The reader will note that (viii) and (xxvi) are identical expressions. Also the appropriate expression for  $V_a$  or  $V_b$  reduces to that of the hypergeometric distribution for the 2-class case since we may write the variance of the unit sample distribution in the form

$$V_1 = \mu_2 - M_1^2.$$

For the 2-class universe  $V_1 = pq$  and (xxi) reduces to

$$V_a = \frac{a(n-a)}{n-1}pq.$$

*Numerical Example.*—From a pack of six cards consisting of the ace, 2, 3, 4, 5 and 6 of clubs, each player takes two cards without replacement, recording as his score the total number of pips. Since  $A$



draws twice, he may select any one of  ${}^6C_2 = 15$  combinations, and  $B$  may choose any one of  ${}^4C_2 = 6$  residual combinations which we set out in the following schema :

<i>A's choice</i>		<i>Possible choice of B</i>				
12	34	35	36	45	46	56
13	24	25	26	45	46	56
14	23	25	26	35	36	56
15	23	24	26	34	36	46
16	23	24	25	34	35	45
23	14	15	16	45	46	56
24	13	15	16	35	36	56
25	13	14	16	34	36	46
26	13	14	15	34	35	45
34	12	15	16	25	26	56
35	12	14	16	24	26	46
36	12	14	15	24	25	45
45	12	13	16	23	26	36
46	12	13	15	23	25	35
56	12	13	14	23	24	43

The corresponding scores of the samples set out in the foregoing schema are as follows :

<i>A's choice</i>		<i>Possible B-scores</i>					
<i>Sample</i>	<i>Score</i>						
12	3	7	8	9	9	10	11
13	4	6	7	8	9	10	11
14	5	5	7	8	8	9	11
15	6	5	6	8	7	9	10
16	7	5	6	7	7	8	9
23	5	5	6	7	9	10	11
24	6	4	6	7	8	9	11
25	7	4	5	7	7	9	10
26	8	4	5	6	7	8	9
34	7	3	6	7	7	8	11
35	8	3	5	7	6	8	10
36	9	3	5	6	6	7	9
45	9	3	4	7	5	8	9
46	10	3	4	6	5	7	8
56	11	3	4	5	5	6	7

Each of the 2-fold samples either  $A$  or  $B$  can draw admits of two permutations. So the number of permutations corresponding to particular scores in this lay-out are in the same ratio as the number of combinations. Consequently, the required frequencies of particular  $A$ -scores associated with particular  $B$ -scores are as exhibited above : and we may summarise the result as a correlation table in which the drift of figures is *downwards* from the top right-hand corner to the left lower corner :

		A									Total	Mean	Variance
		3	4	5	6	7	8	9	10	11			
B	3	..	..	..	..	1	1	2	1	1	6	9.0	$\frac{1.0}{6}$
	4	..	..	..	1	1	1	1	1	1	6	8.5	$\frac{3.5}{12}$
	5	..	..	2	1	2	2	2	1	2	12	8.0	4
	6	..	1	1	2	2	2	2	1	1	12	7.5	$\frac{4.7}{12}$
	7	1	1	2	2	6	2	2	1	1	18	7.0	$\frac{7.0}{18}$
	8	1	1	2	2	2	2	1	1	..	12	6.5	$\frac{4.7}{12}$
	9	2	1	2	2	2	1	2	..	..	12	6.0	4
	10	1	1	1	1	1	1	..	..	..	6	5.5	$\frac{3.5}{12}$
	11	1	1	2	1	1	..	..	..	..	6	5.0	$\frac{1.0}{6}$
	Total	6	6	12	12	18	12	12	6	6	90	7	$\frac{14}{3}$
Mean	9.0	8.5	8.0	7.5	7.0	6.5	6.0	5.5	5.0	7	—	—	
Variance	$\frac{1.0}{6}$	$\frac{3.5}{12}$	4	$\frac{4.7}{12}$	$\frac{7.0}{18}$	$\frac{4.7}{12}$	4	$\frac{3.5}{12}$	$\frac{1.0}{6}$	$\frac{14}{3}$	—	—	

From the data contained in the correlation table we obtain

$$\text{Cov}(x_a, x_b) = -\frac{7}{3}; V_a = \frac{1.4}{3} = V_b;$$

$$\therefore r_{ab} = -\frac{1}{2}.$$

For the inter-class and intra-class variances we have

$$V(M_{ab}) = \frac{7}{3} = V(M_{ba});$$

$$M(V_{ab}) = \frac{7}{3} = M(V_{ba});$$

$$V(M_{ab}) + M(V_{ab}) = \frac{1.4}{3} = V_a;$$

$$V(M_{ba}) + M(V_{ba}) = \frac{1.4}{3} = V_b;$$

$$\therefore V(M_{ba})/V_b = V(M_{ab})/V_a = \frac{1}{4} = r_{ab}^2.$$

The sampling fractions are  $f_a = \frac{1}{3} = f_b$ ,

$$\therefore (f_a \cdot f_b)/(1 - f_a)(1 - f_b) = \frac{1}{4} = r_{ab}^2.$$

### Partition of Variance

Some current text-books suggest to the readers that the square of the product-moment coefficient is a just measure of *explained* variation. We have seen that this is true of correlation in the *consequential* domain of the Umpire Bonus set-up, and regression is always then linear. In the concurrent domain the square of the product moment coefficient is not a true measure of explained variation, though regression *may* in fact be linear. Regression is linear in both dimensions of the non-replacement model; but no meaningful partition of variance is possible. There is a negative correlation between the score of player B and the antecedent score of player A; but the variance of the B-score distribution is exactly the same as it would be if A had not drawn a sample. Evidently, therefore, the square of the product-moment coefficient is not a measure of how much the circumstance of A's antecedent choice contributes to the total variance of the distribution of the B-score.



### EXERCISE 12.02

Set up the correlation table for the long run result and evaluate  $M_a$ ,  $M_b$ ,  $V(M_{b \cdot a})$ ,  $M(V_{b \cdot a})$ ,  $V(M_{a \cdot b})$ ,  $M(V_{a \cdot b})$ ,  $k_{ba}$ ,  $k_{ab}$  and  $r_{ab}$  for the following situations. In Examples 4-8 each player adds the umpire's score to his independent score.

1. A pack consists of 10 cards numbered 1 to 10.  $A$  takes 3 without replacement and retains them.  $B$  takes 4 simultaneously, the players recording their total scores by adding the denominations of the cards.
2. Another pack containing *hearts* only consists of 6 *aces*, 3 *twos*, 4 *threes* and 2 *fives*. Player  $A$  takes 4 without replacement, retaining them. Player  $B$  takes 3.
3. An urn contains 15 balls of which 5 are red and 10 are black. The players draw without replacement,  $A$  taking 2,  $B$  taking 5. They count as their scores the total number of red balls in the sample.
4. A pack consists of 4 cards numbered 1 to 4. The umpire takes 2 without replacement. After noting his score he returns the cards to the pack. The cards are then shuffled, the top two are given to  $A$  and the other two to  $B$ .
5. The situation develops as in 4. After return of the umpire's cards to the pack,  $A$  takes one card and retains it, then  $B$  draws two, without replacement.
6. A pack consists of six cards numbered 1 to 6. The umpire draws two without replacement and retains them.  $A$  then takes two from the remainder and  $B$  gets the two that are left. Does reversing the order of taking the cards make any difference?
7. A pack consists of 9 cards, numbered 1 to 9. The umpire takes 2, without replacement and retains them.  $A$  draws 2, without replacement, from the remaining 7, after noting down his score he returns them to the pack.  $B$  now takes 2 without replacement.
8. Four black and 2 white balls are placed in an urn. The umpire draws out 2 simultaneously, scoring the number of black balls. He replaces them in the urn.  $A$  now draws 2 simultaneously and retains them.  $B$  draws 3 simultaneously from the remainder.

### 12.03 THE TWO-PACK MODEL

We shall now examine a model situation, which involves correlation with linear regression in both dimensions, as in the foregoing section. As is true of the non-replacement model, the causal nexus is not a concurrent relationship like that of the two players of the Umpire Bonus set-up. Nor is the type of reciprocal constraint of a sort which we can rightly describe as consequential in the sense that the score of the player in 12.01 is consequential to that of the umpire.

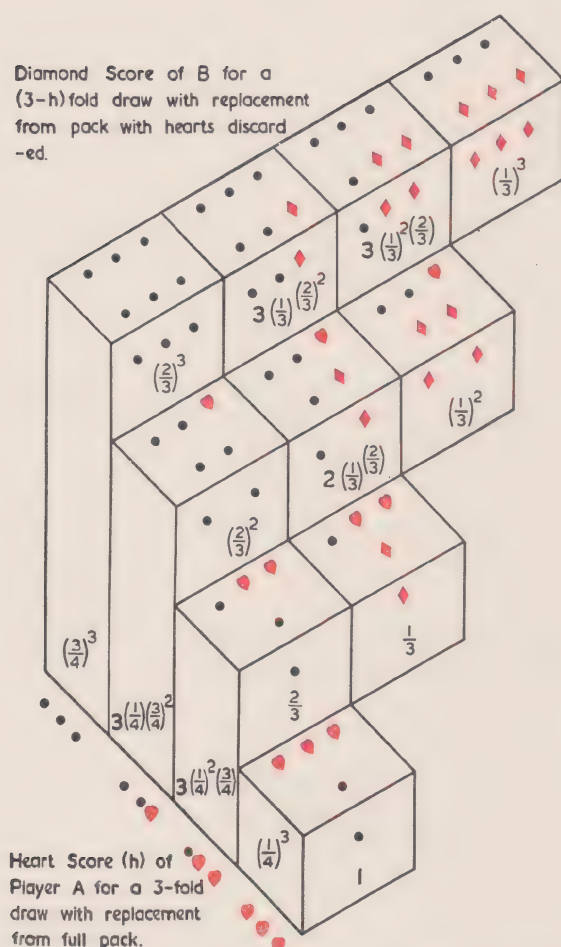
We suppose that two players extract samples *with* replacement, one from a full card pack and one from an otherwise identical pack containing *no* hearts. Player  $A$  takes  $a$  cards from the full pack and records his heart score ( $x_a$ ). Player  $B$  takes  $(a - x_a)$  cards from the pack which contains no hearts and records as his score ( $x_b$ ) the number of diamonds in the sample. We may

denote the numbers of cards in the  $n$ -fold pack from which  $A$  draws as follows :

$$\begin{array}{ccc} \text{Hearts} & \text{Diamonds} & \text{Others} \\ np_a & np_b & np_c \end{array}$$

Evidently  $(np_a + np_b + np_c) = n$  and  $(p_a + p_b + p_c) = 1$ .

### THE TWO-PACK MODEL.



B SCORE

A SCORE	B SCORE			
	0	1	2	3
0	$\begin{array}{c} \bigcirc \times \bigcirc \\ (\frac{3}{4})^3 (\frac{2}{3})^3 \end{array}$	$\begin{array}{c} \bigcirc \times 1 \\ 3 (\frac{3}{4})^3 (\frac{1}{3}) (\frac{2}{3})^2 \end{array}$	$\begin{array}{c} \bigcirc \times 2 \\ 3 (\frac{3}{4})^3 (\frac{1}{3})^2 (\frac{2}{3}) \end{array}$	$\begin{array}{c} \bigcirc \times 3 \\ (\frac{3}{4})^3 (\frac{1}{3})^3 \end{array}$
1	$\begin{array}{c} 1 \times \bigcirc \\ 3 (\frac{1}{4}) (\frac{3}{4})^2 (\frac{2}{3})^2 \end{array}$	$\begin{array}{c} 1 \times 1 \\ 6 (\frac{1}{4}) (\frac{3}{4})^2 (\frac{1}{3}) (\frac{2}{3}) \end{array}$	$\begin{array}{c} 1 \times 2 \\ 3 (\frac{1}{4}) (\frac{3}{4})^2 (\frac{1}{3})^2 \end{array}$	$\begin{array}{c} 1 \times 3 \\ \bigcirc \end{array}$
2	$\begin{array}{c} 2 \times \bigcirc \\ 3 (\frac{1}{4})^2 (\frac{3}{4}) (\frac{2}{3}) \end{array}$	$\begin{array}{c} 2 \times 1 \\ 3 (\frac{1}{4})^2 (\frac{3}{4}) (\frac{1}{3}) \end{array}$	$\begin{array}{c} 2 \times 2 \\ \bigcirc \end{array}$	$\begin{array}{c} 2 \times 3 \\ \bigcirc \end{array}$
3	$\begin{array}{c} 3 \times \bigcirc \\ (\frac{1}{4})^3 \end{array}$	$\begin{array}{c} 3 \times 1 \\ \bigcirc \end{array}$	$\begin{array}{c} 3 \times 2 \\ \bigcirc \end{array}$	$\begin{array}{c} 3 \times 3 \\ \bigcirc \end{array}$

FIG. 92. The Two-Pack Model. Player  $A$  takes 3 cards from one full pack replacing each before drawing another, and recording as his score the number of hearts in the sample. Player  $B$  takes a sample from an otherwise full pack after discarding all hearts, the size of sample being 3, 2, 1 or 0 according as  $A$  scores  $h = 0, 1, 2$  or 3. The score of player  $B$  is the number of diamonds in the  $(3-h)$  fold sample.

In  $B$ 's pack there are  $(n - np_a)$  cards of which  $np_b$  are diamonds. Thus the proportion of diamonds in  $B$ 's pack is  $p_b \div (1 - p_a)$ , whence

$$M_{b.a} = \frac{(a - x_a)p_b}{1 - p_a} \quad \text{and} \quad V_{b.a} = (a - x_a) \left( \frac{p_b}{1 - p_a} \right) \left( \frac{1 - p_a - p_b}{1 - p_a} \right). \quad (i)$$

Evidently,  $M_a = ap_a$ , and  $V_a = ap_a(1 - p_a)$ . Since  $E_a(M_{b.a}) = M_b$ ,

$$M_b = \frac{ap_b}{1 - p_a} - \frac{p_b E_a(x_a)}{1 - p_a} = \frac{ap_b}{1 - p_a} - \frac{M_a \cdot p_b}{1 - p_a},$$

$$\therefore M_{b.a} - M_b = \frac{-p_b}{1 - p_a} (x_a - M_a) \quad (ii)$$



Thus regression of  $x_b$  on  $x_a$  is linear, the regression coefficient being

$$k_{ba} = \frac{-p_b}{1-p_a} \quad \text{. . . . . (iii)}$$

$$\therefore \text{Cov}(x_a, x_b) = k_{ba} \cdot V_a = -a \cdot p_a \cdot p_b \quad \text{. . . . . (iv)}$$

Since regression is linear we may write

$$V(M_{b \cdot a}) = k_{ba}^2 \cdot V_a = \frac{a \cdot p_a \cdot p_b^2}{1-p_a} \quad \text{. . . . . (v)}$$

The variance of  $B$ 's score for a fixed value of  $x_a$  is

$$\begin{aligned} V_{b \cdot a} &= \frac{(a - x_a)p_b(1 - p_a - p_b)}{(1 - p_a)^2}, \\ \therefore M(V_{b \cdot a}) &= \frac{ap_b(1 - p_a - p_b)}{(1 - p_a)^2} - \frac{p_b(1 - p_a - p_b) \cdot ap_a}{(1 - p_a)^2} \\ &= \frac{ap_b(1 - p_a - p_b)}{(1 - p_a)}. \end{aligned}$$

We thus derive

$$\begin{aligned} V_b &= V(M_{b \cdot a}) + M(V_{b \cdot a}) = \frac{ap_ap_b^2}{1-p_a} + \frac{ap_b(1-p_a-p_b)}{1-p_a}, \\ \therefore V_b &= ap_b(1-p_b) \quad \text{. . . . . (vi)} \end{aligned}$$

Hence we have

$$r_{ab} = -\sqrt{\frac{p_ap_b}{(1-p_a)(1-p_b)}} \quad \text{. . . . . (vii)}$$

*Example.*— $A$  draws 3 cards from a full pack, scoring hearts as success.  $B$  draws  $(3 - h)$  cards from a 39-card pack containing 13 diamonds and scores diamonds as success.

	0	1	2	3	Total	Mean	Variance
0	8	12	6	1	27	1	$\frac{6}{9}$
1	12	12	3	0	27	$\frac{2}{3}$	$\frac{4}{9}$
2	6	3	0	0	9	$\frac{1}{3}$	$\frac{2}{9}$
3	1	0	0	0	1	0	$\frac{0}{9}$
Total	27	27	9	1	64	$\frac{3}{4}$	$\frac{9}{16}$
Mean	1	$\frac{2}{3}$	$\frac{1}{3}$	0	$\frac{3}{4}$	...	...
Variance	$\frac{6}{9}$	$\frac{4}{9}$	$\frac{2}{9}$	$\frac{0}{9}$	$\frac{9}{16}$	...	...

$$\begin{aligned} M_a &= \frac{3}{4} = M_b; \quad V_a = \frac{9}{16} = V_b; \\ \text{Cov}(x_a, x_b) &= -\frac{3}{16}; \quad r_{ab}^2 = \frac{1}{9}; \quad k_{ba} = -\frac{1}{3} = k_{ab}; \\ M(V_{b \cdot a}) &= \frac{1}{2} = M(V_{a \cdot b}); \quad V(M_{b \cdot a}) = \frac{1}{16} = V(M_{a \cdot b}); \\ V(M_{b \cdot a})/V_b &= \frac{1}{9} = V(M_{a \cdot b})/V_a. \end{aligned}$$

## EXERCISE 12.03

Investigate the following model situations and construct a grid like that of Fig. 92 for each.

1. Player *A* takes 2 cards with replacement from a full pack recording as his score the total number ( $s$ ) of spades. *B* takes  $(2 - s)$  cards with replacement from a 39-card pack containing no spades, but otherwise complete, and scores diamonds as successes.

2. Player *A* takes *with* replacement 3 balls from an urn containing 5 *red* ones, 7 *black* and 13 *white*, recording as his score the number ( $r$ ) of red balls in the sample. Player *B* takes  $(3-r)$  balls with replacement from an urn containing only 5 *red* and 7 *black*, again scoring red balls.

3. What would be the result if the players of Example 2 did not replace?

## THE RECTANGULAR UNIT SAMPLE MODEL



















Size of Pack ( $x_0$ )		1	2	3	4	5	6		
									
Expected Score ( $x_b$ ) of Unit Sample								Mean ( $M_{ab}$ )	Variance ( $V_{ab}$ )
1		1	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{1}{4}$	$\frac{1}{5}$	$\frac{1}{6}$	2.449	2.574
2		0	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{1}{4}$	$\frac{1}{5}$	$\frac{1}{6}$	3.448	1.902
3		0	0	$\frac{1}{3}$	$\frac{1}{4}$	$\frac{1}{5}$	$\frac{1}{6}$	4.210	1.219
4		0	0	0	$\frac{1}{4}$	$\frac{1}{5}$	$\frac{1}{6}$	4.865	.657
5		0	0	0	0	$\frac{1}{5}$	$\frac{1}{6}$	5.455	.248
6		0	0	0	0	0	$\frac{1}{6}$	6	0
Mean ( $M_{ba}$ )		1	1.5	2	2.5	3	3.5		
Variance ( $V_{ba}$ )		0	$\frac{3}{12}$	$\frac{8}{12}$	$\frac{15}{12}$	$\frac{24}{12}$	$\frac{35}{12}$		

FIG. 93. *Rectangular Unit Sample Model.* The player takes *one* from each of 6 packs respectively containing 1, 2, 3 . . . 6 cards, all clubs numbered consecutively from 1. The marginal entries respectively show size of pack and the player's expectation.



# 12.04 THE UNIT SAMPLE RECTANGULAR MODEL

The two preceding models illustrated situations in which correlation goes with linear regression in both dimensions, though we cannot properly regard the law of association as linear in any other sense. We shall now examine a situation in which correlation goes with linear regression in only one dimension. The rule of the game is as follows: the player chooses one card from each of a set of packs containing different numbers of cards *consecutively* numbered 1, 2, 3 . . . etc., and records as his score the number on the card chosen. In each numerical illustration cited below, the player takes one card from a set of card packs containing 1, 2, 3, up to 6 cards of the club suit starting with the ace. Thus the smallest pack consists of an ace only, the 2-fold pack of the ace and 2, the largest pack containing the ace, 2, 3, 4, 5 and 6 of clubs. The score  $x_a$  is the number of cards per pack, the player's score being  $x_b$  clubs in the unit trial from any particular pack. The maximum score ( $N$ ) of the player is therefore 6.

*Example 1.*—There are 6 packs in all, so that the player takes 1 card from each pack in a single set of trials. Thus the  $A$ -score distribution is rectangular like that of  $x_b$  within a column.

		No. of cards per pack ( $x_a$ )						
		1	2	3	4	5	6	Total
<i>Player's score (<math>x_b</math>)</i>	1	120	60	40	30	24	20	294
	2	—	60	40	30	24	20	174
	3	—	—	40	30	24	20	114
	4	—	—	—	30	24	20	74
	5	—	—	—	—	24	20	44
	6	—	—	—	—	—	20	20
Total		120	120	120	120	120	120	720
Mean		1.0	1.5	2.0	2.5	3.0	3.5	2.25
Variance		0	$\frac{1}{4}$	$\frac{2}{3}$	$\frac{5}{4}$	2.0	$\frac{3.5}{1.2}$	3.5

$$M_a = 3.5; k_{ba} = \frac{1}{2}; M_b = 2.25; V_a = \frac{3.5}{1.2};$$

$$V(M_{b \cdot a}) = \frac{3.5}{4.8}; M(V_{b \cdot a}) = \frac{8.5}{7.2}; V_b = \frac{2.7.5}{1.4.4} = V(M_{b \cdot a}) + M(V_{b \cdot a});$$

$$k_{ba}^2 \cdot V_a = \frac{3.5}{4.8} = V(M_{b \cdot a}); k_{ba} \cdot V_a = \frac{3.5}{2.4} = Cov(x_a, x_b);$$

$$\frac{V(M_{b \cdot a})}{V_b} = \frac{2.1}{5.5} = r_{ab}^2 = \frac{3(N+1)}{7N+13}.$$

*Example 2.*—A set of unit trials involves choice of 1 card from each of 243 packs, the  $A$ -score distribution being defined by successive terms of the binomial  $(\frac{2}{3} + \frac{1}{3})^5$ , i.e. the set-up is

No. of cards per pack	.	1	2	3	4	5	6	
No. of packs	.	.	32	80	80	40	10	1

		No. of cards per pack ( $x_a$ )						
		1	2	3	4	5	6	Total
<i>Player's score (<math>x_b</math>)</i>	1	192	240	160	60	12	1	665
	2	—	240	160	60	12	1	473
	3	—	—	160	60	12	1	233
	4	—	—	—	60	12	1	73
	5	—	—	—	—	12	1	13
	6	—	—	—	—	—	1	1
Total		192	480	480	240	60	6	1458
Mean		1	1.5	2.0	2.5	3.0	3.5	$\frac{11}{6}$

$$k_{ba} = \frac{1}{2}; V_a = \frac{1.0}{9} = (N-1)pq;$$

$$V(M_{b \cdot a}) = \frac{5}{18} = k_{ba}^2 V_a; V_b = \frac{9.5}{108};$$

$$\frac{V(M_{b \cdot a})}{V_b} = \frac{6}{19} = r_{ba}^2 = \frac{3q}{6 + (N-5)p}.$$

\*   \*   \*   \*   \*   \*   \*

Models of this class are not without relevance to practical affairs, since they illustrate the genesis of a coefficient of correlation possibly serviceable as a summarising index of the influence of birth order on the incidence of rare congenital conditions if they occur singly in a sibship. If parity has no effect on the long run expectation w.r.t. birth rank of affected individuals in a sibship  $s$  members will be equal for rank 1, 2 . . .  $s$ . A grid lay-out with family size as one set of border scores and birth rank of affected individuals as the alternate set will then exhibit a triangular contour; and the birth rank mean will increase by equal steps w.r.t. equally spaced family size. On the same assumption there will be a correlation between the two sets of scores; and the product-moment coefficient must be less than unity. To the extent that affected individuals crop up mostly among first births or towards the end of the family, the observed value of the coefficient will approach zero or unity respectively. Thus the discrepancy between the value of the product-moment coefficient ( $r_o$ ) computed from observed data and its value ( $r_{ab}$ ) computed in accordance with the null hypothesis of equal expectation is indicative of the extent to which birth rank influences the occurrence. If  $r_o$  is product-moment coefficient for the *observed* bivariate distribution of affected individuals to each of which we assign one score ( $x_a$ ) in virtue of family size and one score ( $x_b$ ) in virtue of birth rank, and  $r_{ab}$  is the corresponding coefficient computed on the assumption that the expected birth rank of the individual has a rectangular distribution as in models of this class we may define an index ( $B_r$ ) which has the value zero when there is no birth rank effect with limits of  $+1$  and  $-1$  respectively, signifying that all effects are last-born or first-born, *viz.* :

$$B_r = \frac{(r_o - r_{ab})(2r_o \cdot r_{ab} - r_o - r_{ab} + 1)}{r_{ab}(1 - r_{ab})}.$$

For this class of models, we have already defined the score  $x_a$  as the number of cards in one of a set of packs from each of which the player selects a single card. He records the number of pips as his own score  $x_b$  referable to a particular pack the cards of which have 1, 2, 3 . . . pips up to  $s$ , the size of the pack. No more than one card of a particular denomination is present in



the pack, and the denominations are consecutive, hence the unit sample distribution with respect to any pack is rectangular. In the biological illustration  $x_a$  and  $x_b$  respectively correspond to family size and birth rank. The mean value of  $x_b$  associated with  $x_a$  is evidently  $\frac{1}{2}(x_a + 1)$ , so that

$$M_b = \frac{1}{2}E_a(x_a + 1) = \frac{1}{2}(M_a + 1),$$

$$\therefore M_{b..a} - M_b = \frac{1}{2}(x_a - M_a).$$

Thus regression is necessarily linear in the  $B$ -dimension of the grid, irrespective of the distribution of card pack size and  $k_{ba} = \frac{1}{2}$ . In virtue of linear regression

$$k_{ba}^2 \cdot V_a = V(M_{b \cdot a}),$$

$$\therefore V(M_{b..a}) = \frac{1}{4}V_a.$$

From the elementary property of the rectangular distribution,

$$V_{b.a} = \frac{x_a^2 - 1}{12},$$

$$\therefore M(V_{b,a}) = \frac{E_a(x_a^2)}{12} - \frac{1}{12} = \frac{V_a + M_a^2}{12} - \frac{1}{12},$$

[illegible]

[illegible]

[illegible]

The evaluation of  $r_{ab}$  now depends solely on the distribution of card pack—in our biological illustration family size ( $x_a$ ). Two cases admit of easy solution:

(a) rectangular distribution of family size from 1 to  $N$  as in our first numerical example :

$$M_a = \frac{N+1}{2}; \quad M_a^2 = \frac{(N+1)^2}{4}; \quad V_a = \frac{N^2-1}{12};$$

$$\therefore r_{ab}^2 = \frac{3(N+1)}{7N+13} . . . . . (\text{iv})$$

(b) *binomial* distribution of family size as in our second numerical example. We assume that consecutive values of  $x_a$  from 1 to  $N$  occur with frequencies defined by successive terms of the binomial  $(q + p)^{N-1}$ , so that

$$M_a = (N - 1)p + 1 \quad \text{and} \quad V_a = (N - 1)pq,$$

$$\therefore V_b = \frac{(N-1)p}{12}[6 + (N-5)p],$$

$$\therefore r_{ab}^2 = \frac{3q}{6 + (N-5)p} \quad . \quad . \quad . \quad . \quad . \quad (v)$$

The variance of the player's score in the unit sample rectangular model set-up admits of a unique partition in terms of the contribution attributable to the circumstance that the card packs are of unequal size in a set of trials only if we agree on a quite arbitrary prescription of how we propose to eliminate the source of variation. The problem admits of a simple solution if our prescription is to replace each individual card pack of a set of unit trials by one and the same pack of  $M_a$  cards, as is possible only if  $M_a$  is an integer. For the residual (*unexplained*) variance  $V_u$  when we eliminate in this way the source of variation arising from the fact that the card packs are not all of the same size we may then write

$$V_u = \frac{M_a^2 - 1}{12}.$$

Hence from (ii) above, since  $V(M_{b..a}) = \frac{1}{4}V_a$ :

$$V_b = \frac{V_a}{3} + V_u = \frac{4}{3}V(M_{b..a}) + V_u.$$

It is however arguable that we might pool the  $N$  packs of variable size in each  $N$ -fold set of trials and record the result of taking  $N$  unit samples with replacement from the composite pack. Such a procedure admits of no singular solution unless we specify the distribution of card pack size, and hence of the unit sample distribution of the composite pack.

#### EXERCISE 12.04

Examine the joint distribution of the player's score at a single trial and the number of cards in the pack each consisting of cards consecutively numbered from 3 upwards when the distribution of the packs are as follows:

<i>Size of Pack</i>	1	2	3	4	5
<i>No. of Packs</i>					
(a)	1	1	1	1	1
(b)	1	2	3	4	5
(c)	1	4	6	4	1

#### 12.05 LEXIS MODELS

The rule of the following model, so named for reasons mentioned in Chapter 9 of Vol. I, is that the player draws the same number ( $s$ ) of balls from each of  $u$  urns, each containing  $n$  balls of which a certain number ( $x_a$ ) are red, the player's score ( $x_b$ ) at a given trial being the number of red balls in the  $s$ -fold sample. Within this framework, we may distinguish two solutions according as the player does or does *not* replace before drawing another ball. When there is replacement the sample distribution is given by the terms of the binomial

$$(\overline{n - x_a + x_a})^r \div n^r.$$

When there is no replacement the definitive binomial is, of course,

$$(\overline{n - x_a + x_a})^{(r)} \div n^{(r)}.$$



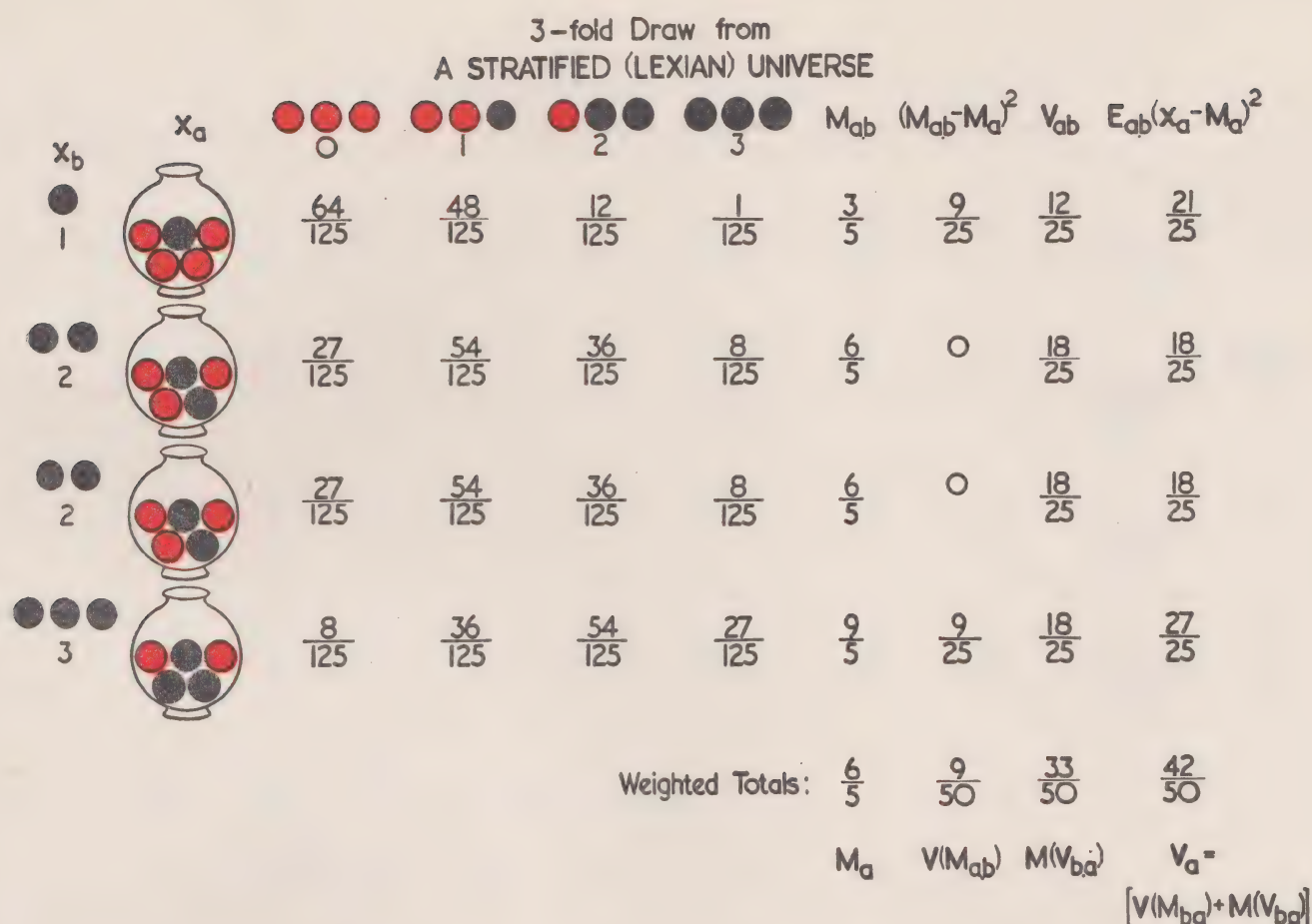


FIG. 94. Correlation in a Lexian Universe. The player draws with replacement 3 balls from each of 4 urns, respectively containing red : black in the ratios 4 : 1, 3 : 2, 3 : 2 and 2 : 3. The player's score is the number of black balls in the sample. The urn score is the number of black in every 5 balls in the urn.

*Example 1.*—Six urns each contain 5 balls of which 0, 1, 2 . . . 5 are respectively red. The player takes 3 balls with replacement from each urn.

		per urn ( $x_a$ )							
Red balls		0	1	2	3	4	5	Total	
per sample ( $x_b$ )	0	125	64	27	8	1	0	225	$\frac{146}{225}$
	1	0	48	54	36	12	0	150	$\frac{312}{150}$
	2	0	12	36	54	48	0	150	$\frac{438}{150}$
	3	0	1	8	27	64	125	225	$\frac{37184}{(225)^2}$
Total		125	125	125	125	125	125	750	$\frac{5}{2}$
$M_{b \cdot a}$		0	$\frac{3}{5}$	$\frac{6}{5}$	$\frac{9}{5}$	$\frac{12}{5}$	3	$\frac{3}{2}$	—
$V_{b \cdot a}$		0	$\frac{12}{25}$	$\frac{18}{25}$	$\frac{18}{25}$	$\frac{12}{25}$	0	$\frac{29}{20}$	—

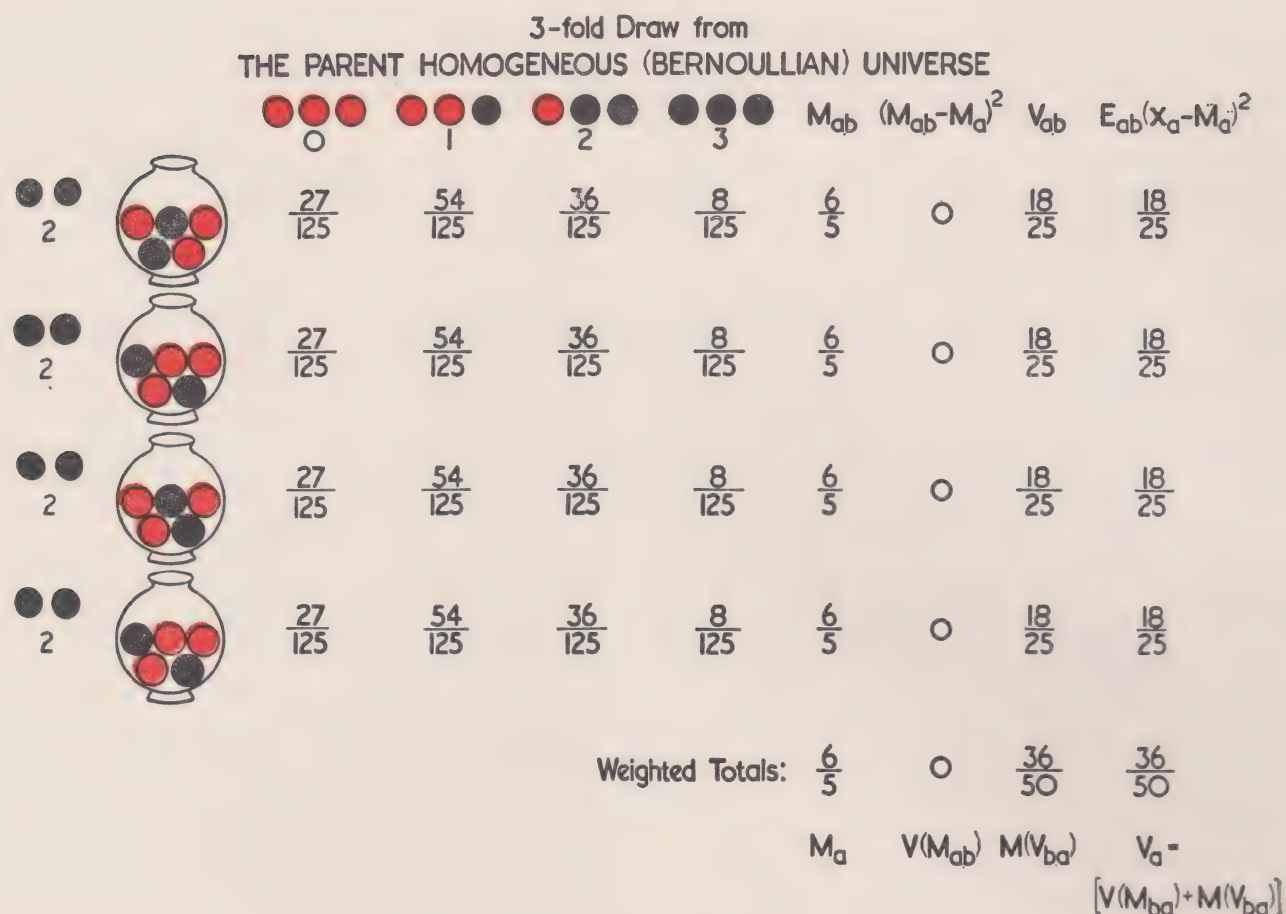


FIG. 95. Homogeneous sampling from 4 urns each equivalent to the pooled contents of the urns in Fig. 94.

$$M_a = \frac{5}{2}; k_{ba} = \frac{3}{5}; M_b = \frac{3}{2};$$

$$V(M_{a \cdot b}) = \frac{717703}{(750)(450)}; M(V_{a \cdot b}) = \frac{66668}{(375)(225)}; V_a = \frac{3 \cdot 5}{1 \cdot 2};$$

$$V(M_{b \cdot a}) = \frac{2 \cdot 1}{3 \cdot 0}; M(V_{b \cdot a}) = \frac{2}{5}; V_b = \frac{2 \cdot 9}{2 \cdot 0};$$

$$\text{Cov}(x_a, x_b) = \frac{7}{4} = k_{ba} \cdot V_a;$$

$$\frac{V(M_{b \cdot a})}{V_b} = r_{ab}^2 = \frac{2 \cdot 1}{2 \cdot 9} = \frac{s(u+1)}{(s-1)(u+1) + 3(2n-u+1)}.$$

*Example 2.*—Six urns each contain 5 balls of which 0, 1, 2 . . . 5 are respectively red. The player may choose *without replacement* 3 balls from each urn.



Red balls	per urn ( $x_a$ )						Total	$M_{a..b}$	$V_{a..b}$
	0	1	2	3	4	5			
0	10	4	1	0	0	0	15	$\frac{2}{5}$	$\frac{2.8}{7.5}$
per 1	0	6	6	3	0	0	15	$\frac{9}{5}$	$\frac{4.2}{7.5}$
sample ( $x_b$ ) 2	0	0	3	6	6	0	15	$\frac{1.6}{5}$	$\frac{4.7}{7.5}$
3	0	0	0	1	4	10	15	$\frac{2.3}{5}$	$\frac{2.8}{7.5}$
Total	10	10	10	10	10	10	60	$\frac{5}{2}$	$\frac{3.5}{1.2}$
$M_{b..a}$	0	$\frac{3}{5}$	$\frac{6}{5}$	$\frac{9}{5}$	$\frac{1.2}{5}$	3	$\frac{3}{2}$	—	—
$V_{b..a}$	0	$\frac{1.2}{5}$	$\frac{1.8}{5}$	$\frac{1.8}{5}$	$\frac{1.2}{5}$	0	$\frac{5}{4}$	—	—

$$M_a = \frac{5}{2}; k_{ba} = \frac{3}{5}; M_b = \frac{3}{2}; k_{ab} = \frac{7}{5};$$

$$V(M_{a..b}) = \frac{4.9}{2.0}; M(V_{a..b}) = \frac{2.8}{6.0}; V(M_{a..b}) + M(V_{a..b}) = \frac{3.5}{1.2} = V_a;$$

$$V(M_{b..a}) = \frac{2.1}{2.0}; M(V_{b..a}) = \frac{1}{5}; V(M_{b..a}) + M(V_{b..a}) = \frac{5}{4} = V_b;$$

$$\text{Cov}(x_a, x_b) = \frac{7}{4} = k_{ba} \cdot V_a; \frac{V(M_{b..a})}{V_b} = \frac{2.1}{5} = r_{ab}^2 = \frac{V(M_{a..b})}{V_a}.$$

*Example 3.*—The player chooses as in Example 1 3 balls from the same 6 different sorts of urns each containing 5 balls, replacing each ball drawn before taking another. There are however 32 urns instead of 6, the  $x_a$  score distribution being binomial in accordance with the terms of  $(\frac{1}{2} + \frac{1}{2})^5$ , i.e. the urns are as follows:

No. of red balls per urn	.	.	0	1	2	3	4	5	
No. of urns	.	.	.	1	5	10	10	5	1

Red balls	per urn ( $x_a$ )						Total	$M_{a..b}$	$V_{a..b}$
	0	1	2	3	4	5			
0	25	64	54	16	1	0	160	$\frac{7}{5}$	$\frac{7.9}{10.0}$
per 1	0	48	108	72	12	0	240	$\frac{1.1}{5}$	$\frac{6.6}{10.0}$
sample ( $x_b$ ) 2	0	12	72	108	48	0	240	$\frac{1.4}{5}$	$\frac{6.6}{10.0}$
3	0	1	16	54	64	25	160	$\frac{1.8}{5}$	$\frac{7.9}{10.0}$
Total	25	125	250	250	125	25	800	—	—
$M_{b..a}$	$\frac{0}{5}$	$\frac{3}{5}$	$\frac{6}{5}$	$\frac{9}{5}$	$\frac{1.2}{1.5}$	$\frac{1.5}{5}$	—	—	—
$V_{b..a}$	0	$\frac{1.2}{2.5}$	$\frac{1.8}{2.5}$	$\frac{1.8}{2.5}$	$\frac{1.2}{2.5}$	0	—	—	—

$$V(M_{a..b}) = \frac{2.6.9}{5.0.0}; M(V_{a..b}) = \frac{8.9}{1.2.5}; V(M_{a..b}) + M(V_{a..b}) = \frac{5}{4} = V_a;$$

$$V(M_{b..a}) = \frac{9}{2.0}; M(V_{b..a}) = \frac{3}{5}; V(M_{b..a}) + M(V_{b..a}) = \frac{2.1}{2.0} = V_b;$$

$$\frac{V(M_{b..a})}{V_b} = r_{ab}^2 = \frac{3}{7} = \frac{s \cdot V_a}{(s-1)V_a + M_a(n - M_a)}.$$

$$r_{ab}^2 = \frac{k_{ba}^2 V_a^2}{V_a \cdot V_b} = \frac{s^2 V_a}{n^2 \cdot V_b}.$$





unique meaning only if we regard the individual urns as sub-samples of a single universe which we can reconstruct on a Bernoullian basis, i.e. as a homogeneous system, by mixing the contents of the  $u$  urns. How much ( $V_E$ ) of the variance of the  $B$ -score distribution then arises from the single relevant circumstance that the universe is *stratified* in virtue of the heterogeneity of the individual urns, we may then assign by deducting what the variance ( $V_U$ ) of the players' score would be if allowed to take  $s$ -fold samples from the reconstituted (Bernoullian) universe of  $un$  balls. Though there is nothing arbitrary about such a definition of *explained* variation, it is important at the outset to recognise that it admits a unique solution only in virtue of an arbitrary property of the model, *viz.* that each sub-universe contains the same number ( $n$ ) of items. With due regard to this limitation we proceed as follows.

If  $f_a$  is the number of urns containing  $x_a$  red balls in the *stratified* universe,  $n_a$  is the total number of red balls and  $M_a$  the mean number per urn :

$$u = \sum_{a=0}^{a=\infty} f_a; \quad n_a = \sum_{a=0}^{a=\infty} f_a \cdot x_a; \quad M_a = \frac{n_a}{u}.$$

The proportion of red balls in the universe as a whole is therefore

$$p_u = \frac{n_a}{un} = \frac{M_a}{n}.$$

If we define the *unexplained* component of variation ( $V_U$ ) as above we may therefore write *with* replacement

$$V_U = sp_u q_u = s \cdot \frac{M_a}{n} \left(1 - \frac{M_a}{n}\right) = \frac{s}{n^2} M_a (n - M_a) \quad . \quad . \quad . \quad (vii)$$

*without* replacement

$$V_U = \frac{s(un - s)}{un - 1} p_u q_u = \frac{s(un - s)}{n^2(un - 1)} \cdot M_a (n - M_a) \quad . \quad . \quad . \quad (viii)$$

In either case by definition

$$V_E = V_b - V_U.$$

Whence from (vii) and (vi) sampling *with* replacement implies

$$V_E = \frac{s(s-1)}{n^2} \cdot V_a.$$

Hence from (ii)

$$V_E = \frac{s-1}{s} \cdot V(M_{b..a}) \quad \text{and} \quad V_U = M(V_{b..a}) + \frac{1}{s} V(M_{b..a}) \quad . \quad . \quad . \quad (ix)$$

When there is *no* replacement, we obtain from (viii) and (iii)

$$\begin{aligned} V_E &= \left[ \frac{s(n-s)}{n^2(n-1)} - \frac{s(un-s)}{n^2(un-1)} \right] M_a (n - M_a) + \frac{s(s-1)}{n(n-1)} V_a \\ &= \frac{n(s-1)}{s(n-1)} V(M_{b..a}) - \frac{s(s-1)(u-1)}{n(n-1)(un-1)} M_a (n - M_a) \quad . \quad . \quad . \quad (x) \end{aligned}$$

The last expression reduces to (ix) when  $n$  is indefinitely large.

In seeking for a partition of variance referable to the source, *viz.* stratification of the universe, we have here explored the result of removing the source of variation by deducting the variance of the player's score in sampling from a parent (Fig. 95) Bernoullian universe, i.e.



a universe constructed by pooling the contents of the urns and extracting  $s$ -fold samples from the composite urn so constituted. Provided  $x_a$  is an exact multiple of  $u$ , we may of course replace the composite urns by  $u$  urns each containing  $n$  balls of which the proportion of red balls is  $p_u$ . When this is so the variance of the player's mean score, i.e.  $V(M_{b..a})$ , is necessarily zero and  $V_b = M(V_{b..a})$ . If the condition of replacement holds good  $V_{b..a}$  does *not* depend on the total number of balls in the urn and it is immaterial whether we define in one or other way stated above what condition we impose when we make the sampling process homogeneous. This is not so when there is *no replacement*. Equations (vii) and (viii) do not then define the alternative stated above, *viz.* sampling from a universe of  $u$  urns each containing  $n$  balls and each having the same proportionate composition as the composite urn containing  $un$  balls. To that extent we may regard the criterion of *explanation* as ambiguous for the case of non-replacement.\*

With one notable exception, it is a common property of the class of models under discussion that regression is always linear in one dimension and in one dimension only. A special case, illustrated by Example 1 above, arises when: (i) we sample without replacement from a stratified universe of which each of  $(n + 1)$  sub-universes contains the same number of items; (ii) the unit trial expectation from one or other sub-universe is  $0, 1, 2 \dots n$ . If the conditions last stated hold good, our numerical example illustrates two properties peculiar to the situation:

- (a) the distribution of the row-means like that of the column border-scores is rectangular:
- (b) regression of the row-scores on the column-scores is linear as is generally true of regression of column-scores on row-scores for other models here dealt with.

Both peculiarities of this set-up are derivable by recourse to properties of figurate numbers dealt with in 11.07. We first recall that

$$\sum_{z=0}^{z=(n-s)} {}^x F_{x+1} \cdot {}^{n-s-z} F_{s-x+1} = {}^{n-s} F_{s+2} = (n+1)_{(s+1)},$$

$$\therefore \sum_{c=0}^{c=n} c_{(x)}(n-c)_{(s-x)} = \frac{(n+1)^{(s+1)}}{(s+1)} \quad \dots \dots \dots \quad \text{(xi)}$$

If the number of columns in the grid is  $(n + 1)$  and the  $A$ -scores run from  $0$  to  $n$  consecutively, the last expression defines the total frequency entries in a single row of the non-replacement Lexian grid in terms of the number of urns  $(n + 1)$  and fixed sample size  $(s)$ . Since this does not depend on the row-score  $(r)$  or the column-score  $(c)$ , (xi) therefore establishes the conclusion that the distribution of row-means is rectangular. The conclusion that the row-means increase as an arithmetic progression follows from the figurate form of the general expression for the row-means, *viz.*:

$$M_{c..r} = \frac{\sum_{c=0}^{c=n} c \cdot c_{(r)}(n-c)_{(s-r)}}{(n+1)_{(s+1)}} \quad \dots \dots \dots \quad \text{(xii)}$$

\* The partition of variance defined by (viii-ix) which is equivalent to the familiar formula for the difference between the variance of the distribution of sampling in the Lexian and Bernoullian universes derives its rationale from postulates totally different from those which justify the identification of explained variation with the so-called *best estimate* of the variance of the column means in a one-way classification of scores of a grid such as one which summarises a plot yield experiment. A Fisher score-grid of this type specifies a single sample from a universe of scores and any parameters based thereon are therefore *estimates*. The Lexian score-frequency grid with which we here deal is a *complete* distribution summarising the results of extracting an indefinitely large number of all possible samples, and the appearance of the factor  $(s - 1) \div s$  on the right hand side of (x) has *no connexion with the problem of estimation*. The equation itself is an exact description of how much variance arises from the circumstance of stratification in the entire universe of choice.

From (xv) of 11.07 we know that

$$\sum_{c=0}^{c=n} c \cdot c_{(r)}(n-c)_{(s-r)} = (r+1)(n+2)_{(s+2)} - (n+1)_{(s+1)}.$$

By substitution in (xii), we obtain

$$M_{c \cdot r} = \frac{(r+1)(n+2)}{(s+2)} - 1 = \frac{r(n+2)}{(s+2)} + \frac{(n-s)}{(s+2)},$$

$$\therefore M_c = E_r(M_{c \cdot r}) = \frac{M_r(n+2)}{(s+2)} + \frac{(n-s)}{(s+2)},$$

$$\therefore M_{c \cdot r} - M_c = \frac{(n+2)}{(s+2)}(r - M_r).$$

Thus there is linear regression of the column-score on the row-score, the regression coefficient being

$$k_{cr} = \frac{n+2}{s+2}.$$

#### SAMPLE SIZE MODEL



#### SAMPLE SCORE ( $x_b$ )

		SAMPLE SCORE ( $x_b$ )						TOTAL	$M_{ab}$
		0	1	2	3	4	5		
SAMPLE SIZE ( $x_a$ )		.....	●	●●	●●●	●●●●	●●●●●		
○	1	768	256	0	0	0	0	1024	$\frac{1}{4}$
○○	2	576	384	64	0	0	0	1024	$\frac{2}{4}$
○○○	3	432	432	144	16	0	0	1024	$\frac{3}{4}$
○○○○	4	324	432	216	48	4	0	1024	$\frac{4}{4}$
○○○○○	5	243	405	270	90	15	1	1024	$\frac{5}{4}$

$$V_a = \frac{11}{16} \quad ; \quad \text{Cov}(x_a, x_b) = \frac{1}{2} \quad ; \quad k_{ba} = \frac{1}{4} \quad ; \quad V_b = 2$$

$$M(V_{ab}) = \frac{9}{16} \quad ; \quad V(M_{ab}) = \frac{1}{8} \quad ; \quad r_{ab}^2 = \frac{2}{11} = r_{ba}^2$$

FIG. 96. Correlation between size of sample and score (number of red balls chosen with replacement) from same urn.



## EXERCISE 12.05

1. Repeat the computations shown above for Examples 1 and 2 when the number of urns is 7 and the number of balls is 0, 1, 2 . . . 6.

2. Repeat the computations for Example 3 when each of 64 urns contains 6 balls as follows :

No. of red balls per urn :	0	1	2	3	4	5	6
No. of urns :	1	6	15	20	15	6	1

## 12.06 SAMPLE SIZE AS A SOURCE OF VARIATION

We shall now consider a set-up in which one score ( $x_a$ ) is the size of the sample the player takes from one and the same urn of  $n$  balls, recording as his score ( $x_b$ ) the number of red ones in the sample. We denote the fixed proportion of red balls in the urn by  $p$ , so that there are  $pn$  red ones in the urn and  $qn = (1 - p)n$  other balls. This prescription admits two variants in virtue of the possibility of imposing the replacement condition or otherwise.

*Example 1.*—The player draws from an urn containing 12 balls of which 8 are red, and separately records the result of taking samples of 1, 2, 3, 4 or 5 balls *without* replacement.

		Sample size ( $x_a$ )					Total	$M_{a..b}$	$V_{a..b}$
		1	2	3	4	5			
Player's score ( $x_b$ )	0	165	45	9	1	0	220	$\frac{13}{10}$	$\frac{351}{1100}$
	1	330	240	108	32	5	715	$\frac{18}{10}$	$\frac{5824}{7150}$
	2	0	210	252	168	70	700	$\frac{157}{50}$	$\frac{2301}{2500}$
	3	0	0	126	224	210	560	$\frac{83}{20}$	$\frac{231}{400}$
	4	0	0	0	70	175	245	$\frac{33}{7}$	$\frac{10}{49}$
	5	0	0	0	0	35	35	5	0
Total		495	495	495	495	495	2475	3	2
$M_{b..a}$		$\frac{2}{3}$	$\frac{4}{3}$	$\frac{6}{3}$	$\frac{8}{3}$	$\frac{10}{3}$	2	—	—
$V_{b..a}$		$\frac{22}{99}$	$\frac{40}{99}$	$\frac{54}{99}$	$\frac{64}{99}$	$\frac{70}{99}$	$\frac{138}{99}$	—	—

$$M_a = 3; k_{ba} = \frac{2}{3}; M_b = 2;$$

$$M(V_{a..b}) = \frac{41757}{61875}; M(V_{b..a}) = \frac{50}{99}; V(M_{a..b}) = \frac{81993}{61875}; V(M_{b..a}) = \frac{8}{9} = k_{b..a}^2 \cdot V_a;$$

$$M(V_{a..b}) + V(M_{a..b}) = 2 = V_a; M(V_{b..a}) + V(M_{b..a}) = \frac{138}{99} = V_b;$$

$$\frac{V(M_{a..b})}{V_a} = \frac{81993}{123750}; \frac{V(M_{b..a})}{V_b} = \frac{44}{69} = r_{ab}^2.$$











## THE ORTHOGONAL LOTTERY

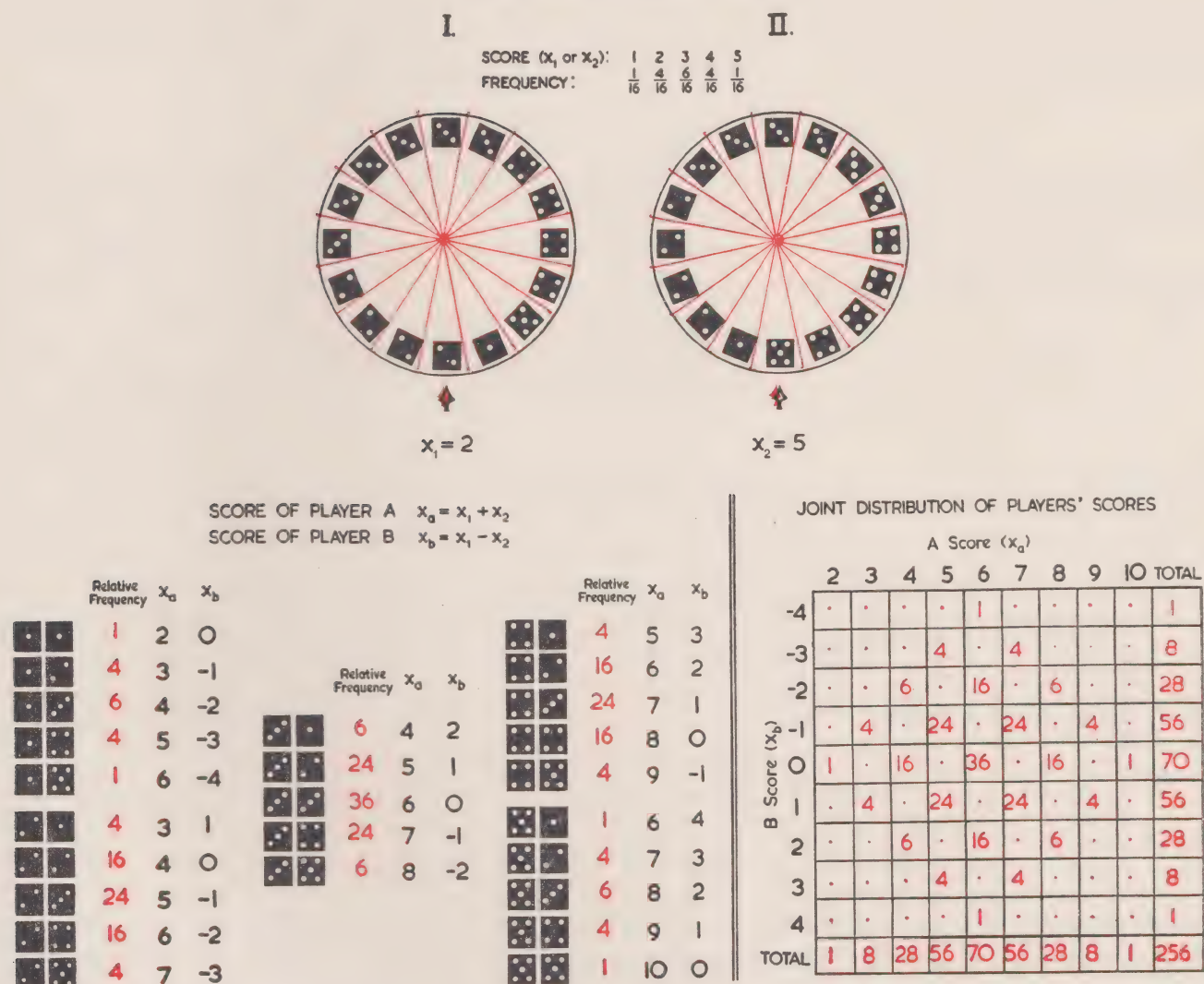


FIG. 97. The Orthogonal Lottery Model. An umpire spins the wheel twice, one player recording the sum as his score, the other the difference.

For the covariance of their scores we have

$$\begin{aligned} Cov(x_a, x_b) &= E(x_a \cdot x_b) - E(x_a)E(x_b) \\ &= E(A_1x_1 + A_2x_2)(B_1x_1 + B_2x_2) - (A_1 + A_2)(B_1 + B_2)M_x^2 \\ &= A_1B_1 E(x_1^2) - A_1B_1M_x^2 + A_2B_2E(x_2^2) - A_2B_2M_x^2 \\ &\quad + (A_1B_2 + A_2B_1)E(x_1, x_2) - (A_1B_2 + A_2B_1)M_x^2 \end{aligned}$$

If we write  $V_x$  for the variance of the single-trial wheel-score distribution, we therefore have

$$Cov(x_a, x_b) = (A_1 B_1 + A_2 B_2) V_x + (A_1 B_2 + A_2 B_1) Cov(x_1, x_2).$$

Since the scores  $x_1$  and  $x_2$  are independent  $Cov(x_1, x_2) = 0$ , and

$$Cov(x_a, x_b) = (A_1B_1 + A_2B_2)V_{x*}$$

The condition of zero covariance is therefore that

[illegible]

$$(A_1^2 + A_2^2 + A_3^2) = (B_1^2 + B_2^2 + B_3^2) = (C_1^2 + C_2^2 + C_3^2) \quad . \quad . \quad . \quad (vi)$$



Similarly for the 2-player set-up with only 2 wheels, it is  $(A_1^2 + A_2^2) = (B_1^2 + B_2^2)$ . The set-up in Fig. 13 is consistent with this condition and with zero covariance since  $(A_1^2 + A_2^2) = 2 = (B_1^2 + B_2^2)$ . When the umpire spins two identical wheels, the joint conditions of unit variance and zero covariance is

$$A_1^2 + A_2^2 = 1 = B_1^2 + B_2^2 \quad \text{and} \quad A_1 B_1 + A_2 B_2 = 0;$$

$$A_1 = \frac{1}{+\sqrt{2}} = A_2 = B_1 \quad \text{and} \quad B_2 = \frac{1}{-\sqrt{2}};$$

$$x_a = \frac{x_1}{\sqrt{2}} + \frac{x_2}{\sqrt{2}} \quad \text{and} \quad x_b = \frac{x_1}{\sqrt{2}} - \frac{x_2}{\sqrt{2}} \quad . \quad . \quad . \quad \text{(vii)}$$

It will be of interest at a later stage to ask what aspect the correlation grid would assume if the number of score classes were so large as to tally closely with the continuous normal distribution. We have (Chapter 3, Vol. I) seen that the binomial  $(\frac{1}{2} + \frac{1}{2})^{20}$  defines a distribution of which this statement is true. Table 1 exhibits the correlation grid of the scores of the players when the universe of scores consists of 21 classes (0 to 20 inclusive) so distributed. The system of scoring is  $x_a = x_1 + x_2$  and  $x_b = x_1 - x_2$  as in Fig. 91.

The ace-contour indicative of zero covariance again asserts itself ; but if our aim is to explore whether zero covariance in a normal universe would conform to the requirements of statistical independence, we are entitled to group our data in conformity with the implicit assumption of continuity as in Table 2.

Each cell of Table 3 exhibits (*above*) the corresponding entry of Table 2 expressed as a proportionate frequency and (*below*) the frequency assigned by the product rule in conformity with the row and column border scores of Table 2. The correspondence is not very good; and it is therefore clear that a bivariate distribution which closely approaches normality in both dimensions does not guarantee independence as a necessary corollary of zero covariance.

### EXERCISE 12.07

The following are the weights ( $A_1, A_a$ , etc. in the first line,  $B_1, B_a$ , etc. in the second, and so on) for the player's score in accordance with the general pattern for player  $K$ :  $x_k = K_1x_1 + K_2x_2 + K_3x_3 \dots$  when  $x_1, x_2$ , etc. represent the score at one trial of each of the roulette wheels with scores (1, 2, 3, 4, 5) and score frequencies (1 : 4 : 6 : 4 : 1) as in Fig. 97:

	Wheel				
Player	1	2	3	4	5
A	1	3	2	4	1
B	4	0	1	1	3
C	1	2	4	2	0
D	2	1	3	1	2
E	0	2	1	1	4

1. Determine the variances of the 5 players' score distributions and those of the first 3 players when they base their scores on those of the first 3 wheels.

TABLE I

$u_2$	0	1	2	3	4	5	6	7	8	9	$u_1$ 10	11	12	13	14	15	16	17	18	19	20	TOTAL
-10										10	1											1
-9											..	10										20
-8									45	..	100	..	45									190
-7								120	..	450	..	450	..	120								1140
-6						210	..	..	1200	..	2025	..	1200	..	210							4845
-5						252	..	2100	..	5400	..	5400	..	2100	..	252						15504
-4					210	..	2520	..	9450	..	14400	..	9450	..	2520	..	210					38760
-3				120	..	2100	..	11340	..	25200	..	25200	..	11340	..	2100	..	120				77520
-2			45	..	1200	..	9450	..	30240	..	44100	..	30240	..	9450	..	1200	..	45			125970
-1		10	..	450	..	5400	..	25200	..	52920	..	52920	..	25200	..	5400	..	450	..	10		167960
0	1	..	100	..	2025	..	14400	..	44100	..	63504	..	44100	..	14400	..	2025	..	100	..	1	184756
1	10	..	..	450	..	5400	..	25200	..	52920	..	52920	..	25200	..	5400	..	450	..	10		167960
2		45	..	..	1200	..	9450	..	30240	..	44100	..	30240	..	9450	..	1200	..	45			125970
3			120	..	..	2100	..	11340	..	25200	..	25200	..	11340	..	2100	..	120				77520
4					210	..	2520	..	9450	..	14400	..	9450	..	2520	..	210					38760
5						252	..	2100	..	5400	..	5400	..	2100	..	252						15504
6							210	..	1200	..	2025	..	1200	..	210							4845
7								120	..	450	..	450	..	120								1140
8									45	..	100	..	45									190
9										10	..	10										20
10											1											1
TOTAL	1	20	190	1140	4845	15504	38760	77520	125970	167960	184756	167960	125970	77520	38760	15504	4845	1140	190	20	1	1048576



TABLE 2

	0-2	3-5	6-8	9-11	12-14	15-17	18-20	TOTAL
(-10)-(-8)	—	—	45	121	45	—	—	211
(-7)-(-5)	—	252	3630	13725	3630	252	—	21489
(-4)-(-2)	45	3630	63000	108900	63000	3630	45	242250
(-1)-(+1)	121	13725	108900	275184	108900	13725	121	520676
(+2)-(+4)	45	3630	63000	108900	63000	3630	45	242250
(+5)-(+7)	—	252	3630	13725	3630	252	—	21489
(+8)-(+10)	—	—	45	121	45	—	—	211
TOTAL	211	21489	242250	520676	242250	21489	211	1048576

TABLE 3

	0-2	3-5	6-8	9-11	12-14	15-17	18-20
(-10)-(-8)	— 0.000	— 0.000	0.000 0.000	0.000 0.000	0.000 0.000	— 0.000	— 0.000
(-7)-(-5)	— 0.000	0.000 0.000	0.003 0.005	0.013 0.010	0.003 0.005	0.000 0.000	— 0.000
(-4)-(-2)	0.000 0.000	0.003 0.005	0.060 0.053	0.104 0.115	0.060 0.053	0.003 0.005	0.000 0.000
(-1)-(+1)	0.000 0.000	0.013 0.010	0.104 0.115	0.262 0.247	0.104 0.115	0.103 0.010	0.000 0.000
(+2)-(+4)	0.000 0.000	0.003 0.005	0.060 0.053	0.104 0.115	0.060 0.053	0.003 0.005	0.000 0.000
(+5)-(+7)	— 0.000	0.000 0.000	0.003 0.005	0.013 0.010	0.003 0.005	0.000 0.000	— 0.000
(+8)-(+10)	— 0.000	— 0.000	0.000 0.000	0.000 0.000	0.000 0.000	— 0.000	— 0.000

2. Make a table of correlations of players' scores of the pattern

	A	B	C	D	E
A	..	..	..	..	..
B	..	..	..	..	..
C	..	..	..	..	..
D	..	..	..	..	..
E	..	..	..	..	..

3. Make a second table showing what fraction of the variance ( $V_{ic}$ ) of the score distribution of each player is attributable to each score component  $K_1x_1$ ,  $K_2x_2$ , etc.

4. Check all the formulae of the preceding section by direct calculation of the above.

5. A lottery wheel like that shown in Fig. 13 has 8 sectors with scores 2 and 8 each in one sector and 4 and 6 each in 3 sectors. Investigate the joint distribution of the players' scores, if they respectively record the score-sum of 2 double spins and the difference between them.

6. A lottery wheel of 32 sectors carries score values 1 or 16 each in 1 sector, 4 or 13 each in 5 sectors and 7 or 10 each in 10 sectors. Investigate the players' joint score distribution if

- (i)  $A$  records the score sum of 2 spins,  $B$  the score difference ;
- (ii)  $A$  records the score sum of 3 spins,  $B$  5 times the score difference of the double spin.

## 12.08 THE STANDARD SCORE SYMBOLISM

We have already used the expression *standard score* or critical ratio in connexion with normal significance tests, its definition being as follows : the standard score is the ratio of the raw-score deviation from the mean to the standard deviation of its distribution. In the theory of the significance test, our assumption is that both the mean and s.d. in this context are their *true*, in contradistinction to their approximate *sample*, values which we commonly use to make the best of a bad job. So defined, the standard mean score of a sample drawn from a normal universe is normally distributed with unit variance ; but the distribution of the approximate standard score then approaches normality only if the sample is large, and its correct distribution is that of the  $t$ -ratio mentioned in Chapter 7 of Vol. I and dealt with in Chapter 16 below. In what follows, the reader may judge from the context whether we employ the term in the exact or approximate sense. Either way, the definition does not presume a normal distribution ; and either way standard scores have certain practical advantages, which commend their use, especially in situations involving the use of correlation and regression formulae. It is for this reason that psychologists employ them extensively.

A special advantage is the simplification of the algebra of correlation, since all such scores have the same mean, i.e. zero, and the same variance, i.e. unity. We shall here use  $z_a$ ,  $z_b$  as the standard scores corresponding to the raw scores  $x_a$  and  $x_b$ , i.e.

$$\frac{X_a}{\sigma_a} = z_a = \frac{x_a - M_a}{\sigma_a} \quad \text{and} \quad \frac{X_b}{\sigma_b} = z_b = \frac{x_b - M_b}{\sigma_b},$$

$$\therefore V_{za} = E\left(\frac{X_a}{\sigma_a}\right)^2 = \frac{1}{\sigma_a^2} E(X_a^2) = \frac{V_a}{\sigma_a^2} = 1, \text{ etc.}$$

By definition we have

$$r_{ab} = \frac{E(X_a \cdot X_b)}{\sigma_a \sigma_b} = E\left(\frac{X_a}{\sigma_a} \cdot \frac{X_b}{\sigma_b}\right) = E(z_a \cdot z_b).$$

Thus the product-moment coefficient of two raw-score distributions is the covariance of the corresponding standard scores. If there is linear concomitant variation, we may write

$$x_a = A_u x_u + A_o x_{a.o} \quad \text{and} \quad X_a = A_u X_u + A_o X_{a.o};$$

$$V_a = A_u^2 V_u + A_o^2 V_{a.o};$$

$$\frac{X_a}{\sigma_a} = \frac{A_u \cdot \sigma_u}{\sigma_a} \cdot \frac{X_u}{\sigma_u} + \frac{A_o \cdot \sigma_{a.o}}{\sigma_a} \cdot \frac{X_{a.o}}{\sigma_{a.o}}.$$



We may then put

$$\frac{A_u \sigma_u}{\sigma_a} = a_u \quad \text{and} \quad \frac{A_o \sigma_{a \cdot o}}{\sigma_a} = a_o;$$

$$\frac{B_u \sigma_u}{\sigma_b} = b_u \quad \text{and} \quad \frac{B_o \sigma_{b \cdot o}}{\sigma_b} = b_o.$$

In standard score form we may then write

$$\begin{aligned} z_a &= a_u z_u + a_o z_{a \cdot o}; \\ z_b &= b_u z_u + b_o z_{b \cdot o}; \\ \therefore V_{za} &= a_u^2 V_{zu} + a_o^2 V_{za \cdot o}, \\ \therefore a_u^2 + a_o^2 &= 1 = b_u^2 + b_o^2. \end{aligned} \quad (i)$$

Also as above

$$\begin{aligned} r_{ab} &= E(z_a \cdot z_b) = a_u b_u E(z_u^2) + a_o b_u E(z_u \cdot z_{a \cdot o}) \\ &\quad + a_u b_o E(z_a \cdot z_{b \cdot o}) + a_o b_o E(z_{a \cdot o} \cdot z_{b \cdot o}) \\ &= a_u b_u V_{zu}. \end{aligned}$$

Since  $V_{zu} = 1$ , we have

$$r_{ab} = a_u b_u = \sqrt{(1 - a_o^2)(1 - b_o^2)} \quad (ii)$$

If regression of  $x_b$  on  $x_a$  is linear we have

$$\begin{aligned} M_{b \cdot a} - M_b &= k_{b \cdot a} \cdot X_a = E_{b \cdot a}(x_b - M_b), \\ \therefore E_{b \cdot a}\left(\frac{X_b}{\sigma_b}\right) &= \frac{k_{ba} \cdot \sigma_a}{\sigma_b} \left(\frac{X_a}{\sigma_a}\right). \end{aligned}$$

Where regression is linear we have seen that

$$k_{ba} = \frac{r_{ab} \cdot \sigma_b}{\sigma_a}.$$

Hence the preceding equation becomes

$$E_{b \cdot a}(z_b) = r_{ab} \cdot z_a \quad (iii)$$

We may express this result as follows : *if regression of the B-score on the A-score is linear, the mean value of standard B-score for a fixed value of the standard A-score is obtained by multiplying the latter by the correlation coefficient.*

To gain familiarity in the use of standard scores, we may here develop the formula for partial correlation cited in 9.04 of Chapter 9, Vol. I on the assumption of *linear concomitant variation*. We assume that there are 2 umpires, and in standard form our definitive equations are

$$\begin{aligned} z_a &= a_u z_u + a_w z_w + a_o z_{a \cdot o}; \quad a_u^2 + a_w^2 + a_o^2 = 1 = V_{za}; \\ z_b &= b_u z_u + b_w z_w + b_o z_{b \cdot o}; \quad b_u^2 + b_w^2 + b_o^2 = 1 = V_{zb}. \end{aligned}$$

In this set-up, the total variances of the players' score distribution are  $V_{za} = 1 = V_{zb}$ . Likewise  $V_u = 1$ , so that

$$\begin{aligned} r_{ab} &= a_u b_u + a_w b_w; \quad r_{au} = a_u; \quad r_{bu} = b_u; \\ \therefore r_{ab} - r_{au} \cdot r_{bu} &= a_w b_w. \end{aligned}$$





To test the validity of the equation of partial correlation in this set-up, we shall require to determine  $r_{bc}$ ,  $r_{ba}$  and  $r_{ca}$  when all three players draw. We already know that

$$r_{ba} = -\sqrt{\frac{ab}{(n-a)(n-b)}} \quad . \quad . \quad . \quad . \quad . \quad (\text{ii})$$

For  $M_a$ ,  $M_b$ ,  $V_a$ ,  $V_b$  we may employ the expressions already derived in 12.02. Since we have shown that the prior choice of  $A$  does not affect the variance or the mean of the  $B$ -score distribution, we may also write

$$M_c = cp \quad \text{and} \quad V_c = \frac{c(n-c)pq}{n-1}. \quad (\text{iii})$$

We may write the mean value of  $x_c$  for a fixed value of  $x_a$  and  $x_b$  as  $M_{c, ba}$ , and its value as

$$M_{c \cdot ba} = \frac{c(np - x_a - x_b)}{n - a - b}.$$

In our grid notation

$$M_{c..a} = E_{b..a}(M_{c..ab}) \quad \text{and} \quad M_{c..b} = E_{a..b}(M_{c..ab}).$$

Whence we derive

[illegible]

Similarly

[illegible]

From (iv) we derive

$$\begin{aligned} E_a(x_a \cdot M_{c \cdot a}) &= \frac{cp^2 \cdot M_a}{n-a} - \frac{c(V_a + M_a^2)}{n-a} = acp^2 - \frac{acpq}{n-1}; \\ \therefore Cov(x_a, x_c) &= -\frac{acpq}{n-1}; \\ \therefore r_{ac} &= -\sqrt{\frac{ac}{(n-a)(n-c)}} . \quad . \quad . \quad . \quad . \quad (vi) \end{aligned}$$

Similarly, we obtain

$$r_{bc} = -\sqrt{\frac{bc}{(n-b)(n-c)}} \quad . \quad . \quad . \quad . \quad . \quad . \quad (\text{vii})$$

From (ii) and (vi) we thus obtain

$$(1 - r_{ab}^2)(1 - r_{ac}^2) = \frac{n^2}{(n - a)^2} \frac{(n - a - b)(n - a - c)}{(n - b)(n - c)},$$

$$r_{ab} \cdot r_{ac} = \frac{a\sqrt{bc}}{(n - a)\sqrt{(n - b)(n - c)}}.$$

Also from (ii)

$$r_{bc} - r_{ab} \cdot r_{ac} = -\frac{n}{n-a} \sqrt{\frac{bc}{(n-b)(n-c)}};$$

$$\therefore \frac{r_{bc} - r_{ab} \cdot r_{ac}}{\sqrt{(1-r_{ab}^2)(1-r_{ac}^2)}} = -\frac{\sqrt{bc}}{\sqrt{(n-a-b)(n-a-c)}}.$$

But we have seen, as in (i) that

$$r_{bc \cdot a} = -\sqrt{\frac{bc}{(n-b)(n-c)}}.$$

Thus the expressions on the right of the last two equations are unequal, i.e. the relation defined by (iv) of 12.08 on the assumption of L.C.V. does not hold good. Nevertheless L.R. continues to apply in both dimensions, since we can write (iv) and (v) in the form

$$M_{c \cdot a} - M_c = \frac{ncp}{n-a} - \frac{c(X_a + ap)}{n-a} - cp;$$

$$M_{c \cdot b} - M_c = \frac{ncp}{n-b} - \frac{c(X_b + bp)}{n-b} - cp;$$

$$\therefore M_{c \cdot a} - M_c = -\frac{c}{n-a} \cdot X_a \quad \text{and} \quad M_{c \cdot b} - M_c = -\frac{c}{n-b} \cdot X_b.$$

What is common ground to L.C.V. and L.R. is that *each suffices to guarantee that the product-moment index has its* essential summarising properties, in that its limits of  $\pm 1$  define perfect correspondence. Its zero value consistent with independence is inherent in its definition, since independence implies zero covariance. That linear regression in one dimension alone suffices to define the limits of  $r_{ab}$  follows from the fact that L.R. implies the identity

$$r_{ab}^2 = \eta_{ba}^2 = \frac{V(M_{b \cdot a})}{V_b} = \frac{V_b - M(V_{b \cdot a})}{V_b}.$$

When correspondence is perfect  $V_{b \cdot a} = 0 = M(V_{b \cdot a})$ , so that  $r_{ab}^2 = 1$ . The summarising properties of the  $p-m$  index when L.C.V. holds good follows from the structure of the 2 types respectively definitive of a consequential and a concurrent relationship. Thus for one common bonus ( $x_u$ )

$$r_{au}^2 = \frac{A_u^2 V_u}{V_a} = \frac{A_u^2 V_u}{A_u^2 V_u + A_o^2 V_{a \cdot o}};$$

$$r_{ab}^2 = \frac{A_u^2 B_u^2 \cdot V_u^2}{V_a V_b} = \frac{A_u^2 B_u^2 V_u^2}{(A_u^2 V_u + A_o^2 V_{a \cdot o})(B_u^2 V_u + B_o^2 V_{b \cdot o})}.$$

Perfect correspondence in this case arises when the player's individual score is zero, so that  $V_{a \cdot o} = 0 = V_{b \cdot o}$ , and both the foregoing expressions reduce to unity.

We may sum up the foregoing discussion :

- (a) linear concomitant variation and linear regression have this in common that each is a sufficient, neither being a necessary, condition of the limiting values set by perfect correspondence to the product-moment index ;



- (b) linear concomitant variation is a sufficient condition of the validity of the equation of partial correlation, but is neither a sufficient nor a necessary condition of linear regression ;
- (c) linear regression is not a sufficient nor a necessary condition either of L.C.V. or of the validity of the equation of partial correlation ;
- (d) L.C.V. defines a causal nexus in the sense that it prescribes two different equations each definitive of a consequential relationship as an adequate and explicit description of a concurrent relationship ;
- (e) L.R. is a descriptive device which implies no causal nexus, since regression may be linear in both dimensions, in one dimension or in neither when the relation between the variates is concurrent.

## CHAPTER 13

# ASSUMPTIONS UNDERLYING ANALYSIS AND SYNTHESIS OF VARIANCE

### 13.00 ANALYSIS OF VARIANCE

No statistical technique has attained in so short a time greater popularity than the one referred to in the title of this chapter ; and it is safe to presume that any reader of this book will have at least a nodding acquaintance with it. One reason for its extensive use in biological and sociological work is the existence of manuals which give very explicit directions for the necessary computations by recourse to examples chosen from contemporary investigations. To the student who understands the logical credentials of the method and the algebraic assumptions of the significance tests it invokes, such instruction is invaluable. Since it is accessible in any well-stocked library, there will be no need to burden ourselves in this context with arithmetical illustrations which the student will find set out in such texts as those of Snedecor, Hagoood, Tippett and others.

In what follows our aim is different. It is all too easy to be led astray by extraneous similarities, if one relies on exemplary material as a guide to the best way of dealing with a statistical problem. Seemingly similar situations in the conduct of enquiries may indeed raise essentially diverse logical issues and may be consistent with very different admissible assumptions about score distributions. Consequently, recourse to a statistical technique without a clear grasp of its rationale is an invitation to its misuse ; and the theme of this chapter is no exception to the rule. For reasons stated in 11.00, we shall not attempt to set forth the appropriate significance tests at this stage. They will be the subject of treatment in Chapter 16. Here our concern is to clarify what we can accomplish by means of the Analysis of Variance, and also what we cannot.

At the outset, it is important to recognise that the term itself covers several statistical procedures of which some have wider applicability than others. Though their several limitations have been the theme of discussion in scientific journals,\* notably by Churchill Eisenhart and by Lee Crump, whose views we quote below, the student who is not a professional mathematician can turn to few, if any, accessible sources from which it is possible to get into focus what factual postulates justify the relevance of the algebraic expressions to a practical situation. Indeed, the intricacy of the relevant computations, so adequately expounded elsewhere, and the novelty of the mathematical technique invoked to justify the appropriate tests of significance, alike conspire to defeat the attempt to do so, unless we distinguish sharply between the following issues :

- (a) the derivation (as in 11.05) of computing devices which rely on tautologies of a grid, and as such have no necessary connexion with the theory of probability ;
- (b) what *causal* assumptions about the real world are implicit in the several procedures we shall discuss in this chapter ;
- (c) what factual assumptions we implicitly make about score distributions in prescribing tests which are the theme of Chapter 16 ;
- (d) the formal algebra (Chapter 15) of the distributions invoked as a basis of such tests.

\* S. Lee Crump (1946), 'Estimation of Variance Components in Analysis of Variance'. *Biometrics* (Amer. Stat. Ass.), **2**, 7.

H. E. Daniels (1939), 'Estimation of Components of Variance'. *J.R.R.S. Supp.*, **6**, 186.

Churchill Eisenhart (1947), 'The Assumptions underlying Analysis of Variance'. *Biometrics*, **3**, 1.



## 13.01 MULTIPLE CRITERIA OF CLASSIFICATION

Statistical treatment of practical issues is possible only when we can *score* our observations in one of two ways: (a) *taxonomically*, as when we say that the number or proportion of items with an attribute  $A$  (e.g. *sickness*, *colour*) is some number  $x$  in a sample of size  $r$ ; (b) *representatively*, when we assign a measurement (e.g. *height*, *weight*) or a number (e.g. *wages*, *seeds per pod*) to each sample item and specify the sample score by a sum, mean or other figure which takes account of their individual values. Having assigned scores of one or other sort to our observations, we may classify them with a view to disclosing some agency which makes them vary. For example, we may divide

- (i) a population of diphtheria patients into those who respectively did and did not receive serum treatment, scoring the sub-samples taxonomically by the proportion of fatal cases in each;
- (ii) a batch of children of the same age and sex into groups respectively travelling at least two miles and less than two miles a day to get to school, scoring the result representatively by recording the median place in the terminal examination or the mean number of absences in a year.

In either case, the practical issue is reducible to the same type of question when stated in the language of statistics used by the predominant school of the last two decades, that of R. A. Fisher. Our first concern is to decide whether the magnitude of the difference observed is consistent with the null hypothesis that the sub-samples come from the same universe. If we may legitimately conclude contrariwise, we may then seek to arrive at some estimate of how the universes themselves differ. If we confine our attention to a single criterion of classification, the effects of treatment or travelling long distances in the foregoing examples, we may often make a two-fold split; but it is sometimes difficult to exclude sources of variation other than those which are our main concern and therefore convenient to divide our material into more than two classes within the framework of the same criterion. One way in which we may then proceed is the theme of what follows.

Let us first be clear about what we mean by *class* and by *criterion* of classification in a situation involving two criteria and two or more classes with respect to each criterion. We shall suppose that we have before us figures w.r.t. a single determination of the red blood cell count of one male and one female of the Angora, Blue Bevan and Polish Giant breeds of rabbits. We may then lay out our scores ( $x_{ij}$ ) gridwise as below.

TABLE I

	Angora	Blue Bevan	Polish Giant	Row
Male	$x_{11}$	$x_{21}$	$x_{31}$	$j = 1$
Female	$x_{12}$	$x_{22}$	$x_{32}$	$j = 2$
Column	$i = 1$	$i = 2$	$i = 3$	

Here we have two criteria of classification *sex* and *breed*, involving 2 classes w.r.t. the first and 3 classes w.r.t. the second. It may well be that sex affects the score value in the sense that the *mean* score of one sex differs from that of the other in the absence of any other *source of variation*, i.e. agency responsible for score differences. To say that sex is the *only* source of variation so defined would signify that the scores of all males are the same and those of all females are the same,



the single male score value (which is then the male *mean* score) being different from the single female score value (which is then the female *mean* score). Thus the variance ( $V_r$ ) of the score distribution within each row will be zero, so that  $M(V_r) = 0$ . Hence the variance of the row mean scores  $V(M_r)$  is equal to the total variance ( $V$ ) in virtue of the tautology of the score grid,  $V = M(V_r) + V(M_r)$ . Contrariwise, the column means will be identical so that  $V(M_c) = 0$  and  $V = M(V_c)$ . By the same token, we might write  $V = V(M_c)$  and  $M(V_c) = 0$  if the only source of variation were the breed, in which case also  $V(M_r) = 0$  and  $M(V_r) = V$ .

In general, neither proposition last stated will be true. For we know that any single determination of the r.b.c. is subject to error of observation and to individual circumstances unconnected with either sex or breed. In virtue of these *residual* sources of variation, the scores within a row or column will differ in the absence of any row-effect attributable to sex or column-effect attributable to breed. Because of the residual effect alone, both row means and column means may therefore differ. If there is no row-effect, the fact that  $V(M_r)$  exceeds zero is then attributable to the residual source of variation alone, as is the fact that  $V(M_c)$  exceeds zero when there is no column-effect. If there is neither a row- nor a column-effect, the values of both  $V(M_r)$  and  $V(M_c)$  depend uniquely on the residual source, and we may expect to discover some relation between them, consistent with that assumption.

To say that there is no sex-effect in this context is to say that row samples come from one and the same universe; and we express this alternatively by saying that the universe is *homogeneous* w.r.t. the row-criterion of classification. To say that there is no breed-effect is to say that column samples come from the same universe and that such a universe is homogeneous w.r.t. the column-criterion. To say that the universe is homogeneous in both dimensions is the same as saying that the residual sources suffice to account for *all* variation.

One class of procedure subsumed under the term *analysis of variance* has as its aim to decide whether a system of scores is homogeneous w.r.t. one or more criteria of classification, i.e. to test the null hypothesis that one or other putative source of variation defined by the classificatory set-up is negligible. Thus the null hypothesis is that sex or breed or both do not significantly affect the r.b.c. in the example last cited. If our 2-way table for 2 criteria of classification has  $c$  columns,  $r$  rows and hence  $rc$  cells in all, homogeneity w.r.t. both criteria signifies that each row-set is a  $c$ -fold, each column-set is an  $r$ -fold and the entire set of scores is an  $rc$ -fold sample from one and the same universe. If so, our problem is:

- (a) to define *consistent* relations between parameters of the score distribution referable to either dimension alone and to the grid as a whole;
- (b) to test the consistency of such parameters.

Any significance test specified by (b) must naturally rely on certain assumptions about the score distribution of the putative common universe. Such assumptions may be more or less plausible in a given situation; but we can define criteria of homogeneity in the sense implied by (a) without invoking them. It will therefore be convenient to reserve discussion of (b) till a later stage.

A corresponding dichotomy is helpful, when we turn to other classes of problems to which the expression analysis of variance may refer. If the universe is not homogeneous, we are entitled to ask how much of the variance is attributable to one or other source. The end in view may then be

- (i) to make an *exhaustive* balance sheet exhibiting what fraction of the total variance arises from each source;
- (ii) more modestly, to assess the *residual* component with a view to specifying the sampling variance of a set of class means in the absence of variation arising from other sources.



More explicitly, (i) signifies a specification of the extent to which we should proportionately reduce the variance of the parent universe, if we eliminated one or other source of variation, e.g. by replacing all males by females or *vice versa* in the foregoing example. It is important to notice that we can reduce it in two ways, and we have at this stage no reason to believe that the results would be the same. They are indeed the same only if we make certain assumptions which have no relevance to the discussion of homogeneity.

To appreciate the meaning of (ii) above, we should recall that the observed variance of the column (*breed*) samples of our foregoing illustration will partly depend on sex if sex is a true source of variation. If our aim is to ascertain the variance of each column mean regarded as a parameter of the breed effect alone, i.e. the true variance of the mean of males alone or of females alone for each breed, our only concern is therefore with one component (the *residual*) of variance.

To construct an exhaustive balance sheet of all the components of variance or to specify a particular component, we have to rely on certain assumptions about the ways in which the several sources of variation contribute to the individual score value; and this raises an issue which does not arise when our sole aim is to test the null hypothesis that one or other source of putative variation is negligible. Needless to say, we can ask no more of our sample than an *estimate* of any component of variance. There then arises the question: to what sampling error are our estimates subject, i.e. what are their fiducial or confidence boundaries? Here we must introduce other assumptions concerning the distribution of the score components. We shall find it easier to steer a way through a labyrinth of practical difficulties, if we examine the simpler issue: how is it possible to construct the balance sheet? Our first task will be an attempt to visualise random distributions of samples classified w.r.t. one, two or three criteria by recourse to statistical models. It will then be possible to exhibit the formal logic of analysis of variance by recourse to the symbolism of 11.05 without the danger of losing ourselves in a maze of symbols.

### 13.02 THE COMPLETE SAMPLING DISTRIBUTION

In Vol. I we have become familiar with the chessboard device as a means of setting out the distribution of variously classified *r*-fold samples from a static *n*-fold universe in conformity with the principle of equipartition of opportunity for association; but we have also acquainted ourselves with the advantage of a somewhat different approach in 12.00, where we have spoken of a *universe in action*, i.e. as a random distribution of all possible *r*-fold samples. In the entire assemblage of samples constituting such a random distribution every item (*score value*) is necessarily present with the same proportionate frequency as in the parent universe. This is sufficiently evident from the build up of the 3-fold toss of a penny when we score heads as 1 and tails as 0:

	0	1		0.0	0.1	1.0	1.1
0	0.0	0.1	0	0.0.0	0.0.1	0.1.0	0.1.1
1	1.0	1.1	1	1.0.0	1.0.1	1.1.0	1.1.1

Two-fold Toss				Three-fold Toss			
Heads 4				Heads 12			
Tails 4				Tails 12			

	1	2	2	3
1	1.1	1.2	1.2	1.3
2	2.1	2.2	2.2	2.3
2	2.1	2.2	2.2	2.3
3	3.1	3.2	3.2	3.3

Numbers of each	.	.	.	8	16	8
-----------------	---	---	---	---	----	---

- (a) the unit-sample variance ( $V_u$ ), being the variance of the distribution of individual scores in the parent universe ;
- (b) the variance ( $r \cdot V_u$ ) of the distribution of the *sum\** of individual score values in  $r$ -fold samples drawn independently from the parent universe with replacement ;
- (c) the variance  $V(M_{r \cdot s}) = (V_u \div r)$  of the distribution of the mean ( $M_{r \cdot s}$ ) of individual score values in such  $r$ -fold samples ;
- (d) the *observed* variance  $V_{r \cdot s}$  of the distribution of individual score values ( $x_j$ ) whose mean value is  $M_{r \cdot s}$  within a *particular* sample, so that

$$V_{r \cdot s} = \sum_{j=1}^{j=r} \frac{(x_j - M_{r \cdot s})^2}{r} \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad (i)$$

\*NOTE—In the domain of *binary* taxonomic scoring, the unit scores are respectively 0 and 1 and the terms of the binomial  $(q + p)^1$  define the unit sampling distribution with mean  $p$  and variance  $V_u = pq$ . The  $r$ -fold score sum is then the  $r$ -fold sample raw score  $(0, 1 \dots r)$  whose distribution corresponds with successive terms of the binomial  $(q + p)^r$  with mean  $rp$  and variance  $rqp$ . The  $r$ -fold sample mean score is the proportionate score whose distribution likewise accords with successive terms of  $(q + p)^r$  for score values  $0, \frac{1}{r}, \frac{2}{r} \dots 1$ , the mean of this distribution being  $p$ .



In accordance with the notation of 11.05, we may write (i) more economically in the form

$$E_r(x_r - M_{r.s})^2 = V_{r.s} = E_r(x_r^2) - M_{r.s}^2 \quad (ii)$$

In Chapter 7 of Vol. I we have seen that the expected (*mean*) value of the sample variance ( $V_{r.s}$ ) is somewhat smaller than that of the unit sample distribution, being in fact defined by the relation

$$M(V_{r.s}) = \frac{r-1}{r} V_u, \quad (iii)$$

We shall now derive this relation by recourse to a notation which makes the meaning of every step explicit. Let us first be clear about what we call the mean sample variance in this context. The statistic ( $V_{r.s}$ ) defined by (i)-(ii) is an  $r$ -fold *sample* statistic. The statistic defined by (iii) is the mean value of this sample statistic in the universe of all  $r$ -fold samples with relative frequencies assigned by successive application of the product rule. We may make the distinction clear by: (a) using  $E_s$  for the operation of extracting the universe mean of the random distribution of any sample parameter; (b) employing the dot notation to distinguish any parameter of the  $r$ -fold sample (e.g.  $M_{r.s}$  or  $V_{r.s}$  for the mean score or variance) from a parameter of the universe (e.g.  $M_r$  or  $V_r$ ). With these conventions we may write

$$M(V_{r.s}) = E_s(V_{r.s}) = E_s[E_r(x_r^2) - M_{r.s}^2] = E_s \cdot E_r(x_r^2) - E_s(M_{r.s}^2);$$

$$V(M_{r.s}) = E_s(M_{r.s} - M_r)^2 = E_s(M_{r.s}^2) - M_r^2;$$

$$M(V_{r.s}) + V(M_{r.s}) = E_s \cdot E_r(x_r^2) - M_r^2 = E_s \cdot E_r(x_r - M_r)^2.$$

The expression on the right of the last equation is the mean value of the square of individual score deviations from the true mean in the universe of all  $r$ -fold samples, i.e. that of the unit sample distribution, so that

$$M(V_{r.s}) + V(M_{r.s}) = V_u.$$

We have also seen that

$$V(M_{r.s}) = \frac{V_u}{r},$$

$$\therefore M(V_{r.s}) = V_u - \frac{V_u}{r} = \frac{r-1}{r} V_u \quad (iv)$$

We may conveniently visualise a complete sample distribution involving one criterion of classification by setting out each permutation of individual scores in the chessboard cell entries as rows of a 2-dimensional score grid. Fig. 98 and Table 2 show the relevant calculations for a score grid exhibiting the random distribution of 3-fold samples of a flat circular die with 1 pip on one face and 2 pips on the other. The binomial  $(\frac{1}{2} + \frac{1}{2})^3$  then defines the unit sample distribution, so that  $M_u = \frac{3}{2}$  and  $V_u = \frac{1}{4}$ . Alternatively, we may lay out the distribution more

and the variance  $pq \div r$  in accordance with (c) above. On this understanding, we may regard taxonomic scoring in the binary universe as a special case of representative scoring.

More generally, we speak of the unit sample distribution as a *binomial variate*, if the frequencies of individual scores correspond to successive terms of  $(q + p)^a$  in the range  $m$  to  $m + a\Delta x$ , when the frequency of a score  $m + x\Delta x$  is given by  $a_{(x)} \cdot p^x \cdot q^{a-x}$ . The mean ( $M_u$ ) of the unit sample distribution is then  $m + ap\Delta x$  and the variance  $apq(\Delta x)^2$ . What we call taxonomic scoring in the binary universe signifies that  $a = 1$ ,  $m = 0$ ,  $\Delta x = 1$  so that  $M_u = p$ . The binary universe of the flat circular die with 1 pip on one face and 2 on the other does not fulfil this specification, since  $\Delta x = 1$ , but  $m = 1$  and  $M_u = (p + 1)$ , though  $V_u = pq$  as when the method of scoring is taxonomic. When  $p = \frac{1}{2} = q$  and  $a = 1$ , the distinction between a rectangular and binomial variate breaks down.







We can thus define a sample statistic whose expected value is the true variance of the universe by the relations

$$\frac{r}{r-1} V_{r.s} = s_r^2 = \frac{\sum_{j=1}^r (x_j - M_{r.s})^2}{r-1} \quad \text{and} \quad E(s_r^2) = V_u \quad (vi)$$

It is customary to speak of  $s_r^2$  so defined as the unbiased estimate of the u.s.d. variance, i.e. variance of the distribution of individual scores in the parent universe.

The foregoing analysis involves classification of the sample by *one criterion*. We can introduce a second criterion of classification, if we toss more than one die. We then have to represent each class of samples by a two-dimensional lay-out, and the entire assemblage of the random sample distribution as a 3-dimensional grid of which each *layer* is a sample. The number of layers having one and the same set of scores in corresponding cells must then tally with the relative frequency of the sample so defined.

Fig. 99 shows the build up of the grid in accordance with the chessboard principle for the 2-fold toss of each of 2 unbiased pennies. If we assume that each toss is a fair toss, the distribution of scores in each *pillar*, each *row-slab* and each *column-slab* will accord with one and the

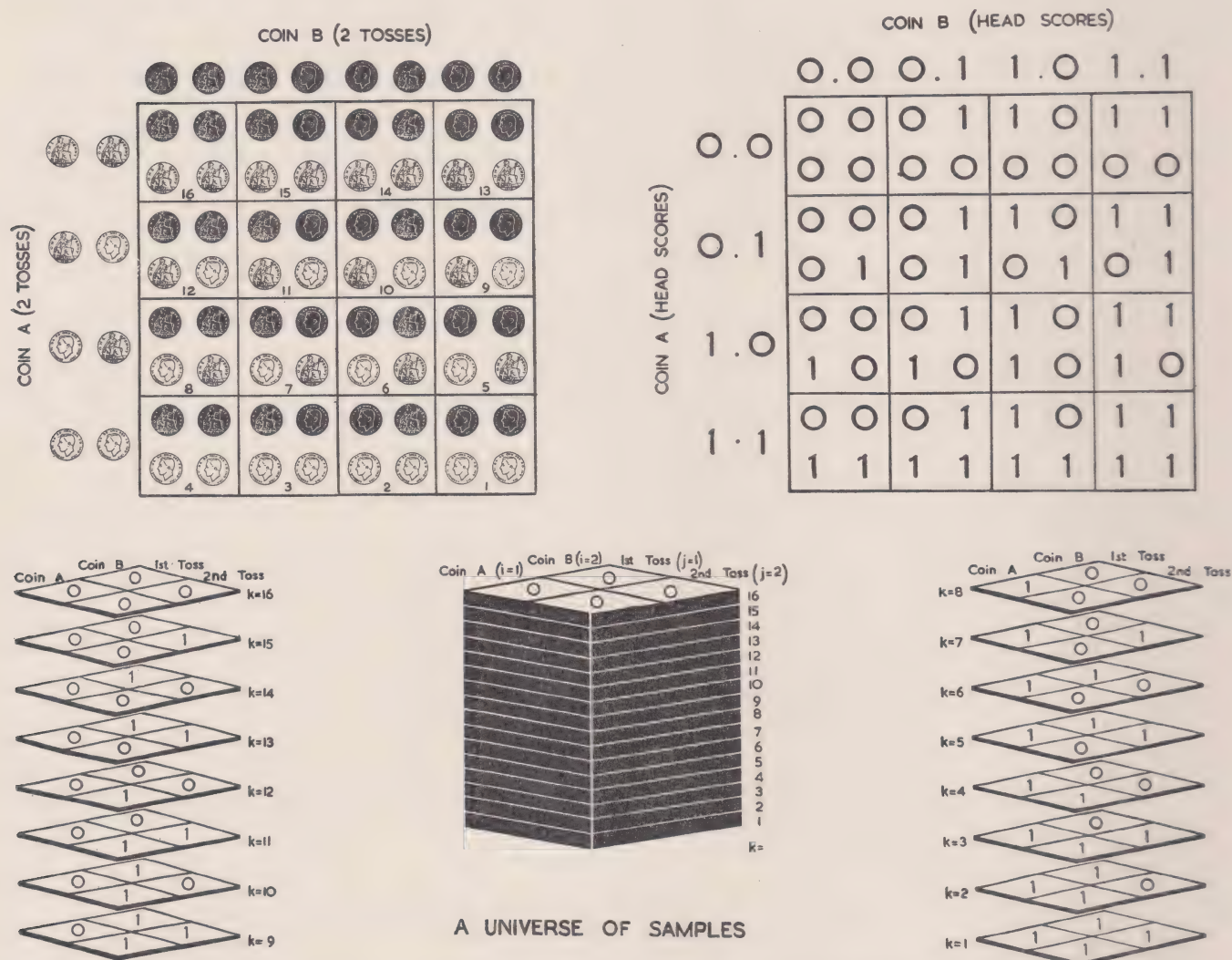


FIG. 99. The results of 2 tosses of an unbiased coin set out as a 2-way classification, the criteria being identity of coin and order of toss. All possible samples are shown in their correct proportions.



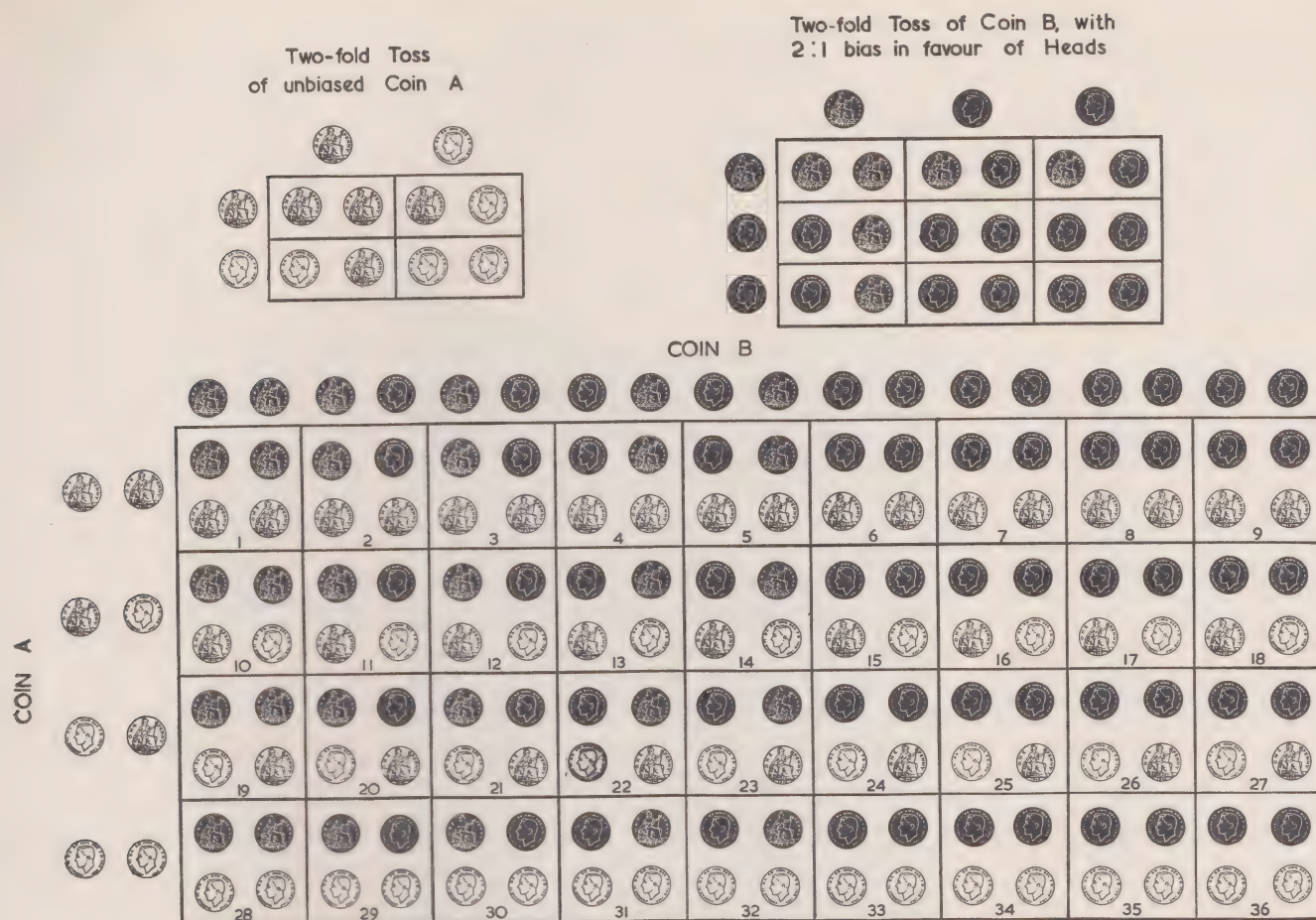


FIG. 100. Chessboard lay-out for 2-fold toss of a biased coin (2 : 1 in favour of heads) exhibiting results for the same 2 criteria of classification as in Fig. 99.

same unit sample distribution, i.e. that of individual scores in the entire 3-dimensional universe of samples. We then say that the universe is *homogeneous for both criteria of classification*. The pillar-mean scores of this example will thus be

		Coin	
		A	B
Toss	1	$\frac{1}{2}$	$\frac{1}{2}$
	2	$\frac{1}{2}$	$\frac{1}{2}$

In this case, each of 16 possible samples of different structure occurs with equal frequency. If one coin has a bias the number of different sample classes is still sixteen ; but their frequencies are not all equal. Fig. 100 shows each stage of the lay-out for the 2-fold toss of two coins, one of which (A) has no bias, while the other (B) falls head upwards twice as often as tails. Thus the definitive binomials of the unit sample distributions are respectively  $(\frac{1}{2} + \frac{1}{2})^1$  and  $(\frac{1}{3} + \frac{2}{3})^1$ .

Here the entries of the same column refer to the same coin, entries of the same row to the same toss-order. Each column-slab in the universe of Fig. 101 thus constitutes a homogeneous sub-universe in the sense that the distribution of individual scores within pillars of one and the same column-slab are identical, but the distributions of scores in different column-slabs are *not* identical. The within-pillar distributions of the same row are not identical ; but the

**SAMPLE DISTRIBUTION**  
 for 2-fold Toss of unbiased Coin A  
 and 2-fold Toss of biased (2:1) Coin B

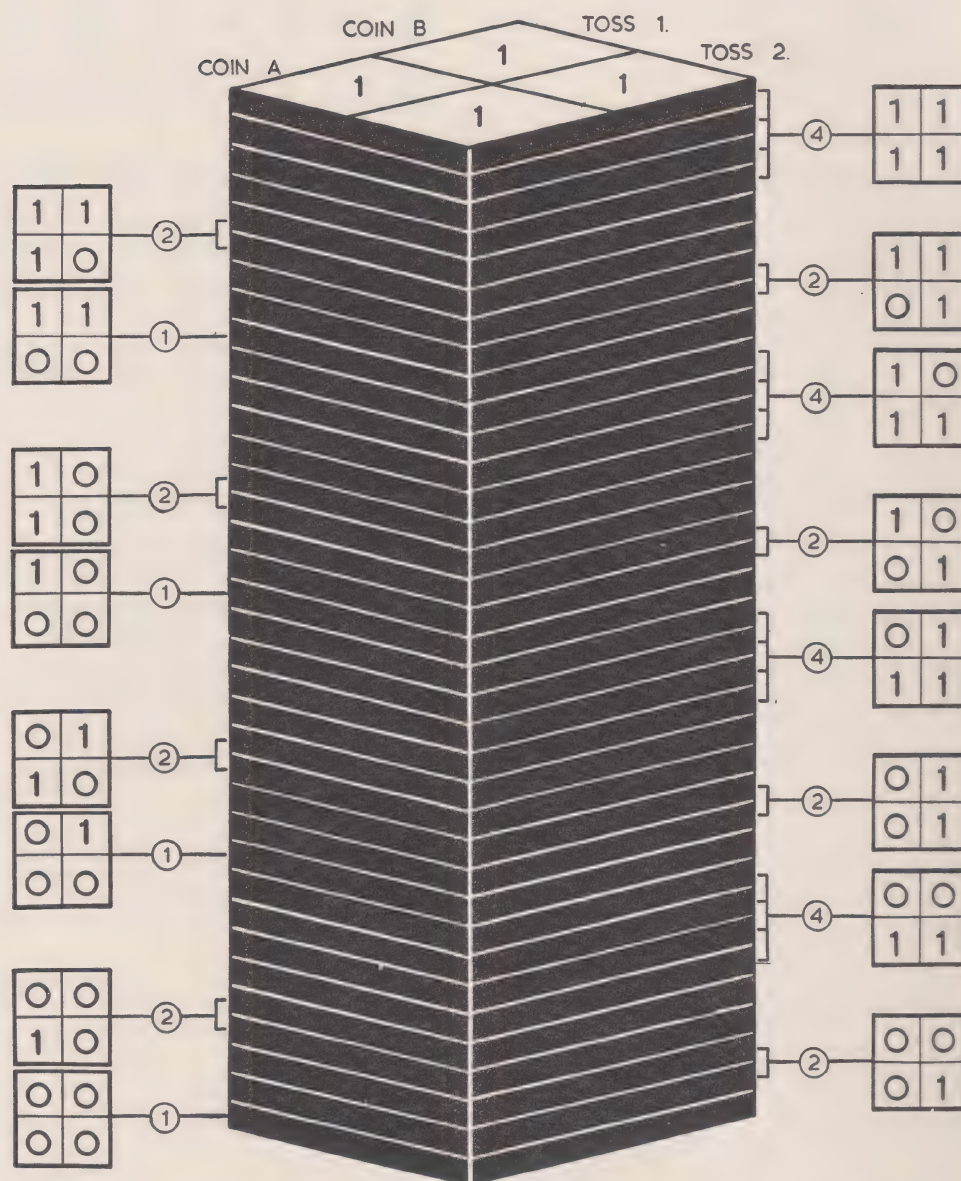


FIG. 101. The 3-dimensional universe of all samples shown in Fig. 100.

row-slab distributions are identical with one another and therefore with that of the universe as a whole. The pillar-mean scores will be

		Coin	
		A	B
Toss	1	$\frac{1}{2}$	$\frac{2}{3}$
	2	$\frac{1}{2}$	$\frac{2}{3}$



Of this set-up, we say that the universe is homogeneous only for the row-criterion of classification. Thus we can define a universe as homogeneous w.r.t. either or both of two criteria of classification, if we conceive it as a complete random distribution of  $rc$ -fold samples classified with respect to one (*column*) criterion involving  $c$  classes and a second (*row*) criterion involving  $r$  classes. So conceived our criteria of homogeneity are as in Table 4.

TABLE 4

HOMOGENEOUS WITH RESPECT TO :	DISTRIBUTION OF INDIVIDUAL SCORES IN			
	<i>Row-Slabs</i>	<i>Column-Slabs</i>	<i>Pillars within row-slab</i>	<i>Pillars within column-slab</i>
<i>Both criteria</i> <i>Row-criterion only</i> <i>Column-criterion only</i>	as for whole grid as for whole grid different	as for whole grid different as for whole grid	as for whole grid different as for row-slab	as for whole grid as for column-slab different

We can still visualise the distribution of samples from a universe classified w.r.t. 3 criteria, if we conceive each sample as a 3-dimensional grid constituting a *stratum* and the universe of samples as a vertical succession of such strata. As a simple illustration we may consider the result of the two-fold toss of 4 coins :

*A* French silver                      *C* American silver  
*B* French copper                      *D* American copper

We have now 3 criteria of classification : (a) toss-order ; (b) metal ; (c) nationality ; and we are entitled to ask whether the universe is homogeneous with reference to any or all, i.e. whether each toss is a fair toss, whether French coins have the same bias (if any) as American coins or whether copper coins have the same bias (if any) as silver coins. We may lay out in one layer of the sample stratum (Fig. 102) French coins by toss-order and metal as our criteria and in the other American coins by toss-order and metal in corresponding dimensions. In this case each stratum consists of 2 layers each of 2 columns and 2 rows. If there are  $n$  classes with respect to the layer-to-layer criterion of classification, the stratum will contain  $ncr$  cells, and the universe grid of  $s$ -strata will consist of  $ncrs$  cells.

Here, as elsewhere, we use the term sample frequency ( $y_s$ ) for the *proportionate* contribution of a sample class to the entire distribution of such classes, so that the sum of all frequencies is unity. When we speak of *relative* frequencies ( $f_s$ ) we signify corresponding *whole* numbers in the same ratio and define  $s$  as their sum, so that  $y_s = (f_s \div s)$ . If the grid representative of a complete sample distribution consists of  $s$  layers (or strata, as the case may be), we assume that there are  $f_s$  layers (or strata) exhibiting the lay-out of a particular sample structure of relative frequency  $f_s$ . We may then define the operation of taking the mean of a sample parameter ( $U_s$ ) alternatively as

$$\sum_0^{\infty} y_s \cdot U_s = E(U_s) = \frac{1}{s} \sum_{k=1}^{k=s} (U_k).$$

If we care to regard the theoretical sampling distribution as continuous, we must interpret the summation as an integral. This does not affect the ensuing argument.

### 13.03 CRITERIA OF HOMOGENEITY

We have already defined in general terms what we mean by criteria of homogeneity in this context, *viz.* the definition of sample parameters whose expected values should be consistent. In what follows, we shall first explore the issue *vis-à-vis* 2 criteria of classification.

[illegible]





The following numerical example referable to a 3 by 3 grid illustrates the computation.

(a)					(b)			
Scores			Total	$T_j^2$	Square Scores			
3	1	8	12	144	9	1	64	
9	7	5	21	441	81	49	25	
6	4	2	12	144	36	16	4	
Total	18	12	15	45	Total			285
$T_i^2$	324	144	225	693				

$$S_a = 285; S = \frac{1}{9}(45^2) = 225; S_c = \frac{6 \cdot 9 \cdot 3}{3} = 231;$$

$$S_r = \frac{7 \cdot 2 \cdot 9}{3} = 243.$$

Estimate based on	Sum of Squares see (x)-(xiv) in 11.05	Divisor	Estimate
Rows	$243 - 225 = 18$	2	9
Columns	$231 - 225 = 6$	2	3
Residual	$285 + 225 - 243 - 231 = 36$	4	9
Total 60			

When our concern is with 3 criteria of classification, the number of possibilities we may wish to explore involve not only the existence of an additional systematic source of variation but every possible *interaction* between all three of them. Which statistic we chose to compare with the residual as a criterion of homogeneity or non-interaction is an issue which will raise less difficulty for the beginner if we defer it till we have examined the balance sheet of 13.06. In the derivation of (i) above we have used the relation

$$E_s(V_{x \cdot s}) = \frac{(rc - 1)}{rc} \sigma^2.$$

Accordingly, we may define a statistic which takes into account variation in both dimensions of the grid by

$$s_t^2 = \frac{rc}{rc - 1} V_{x \cdot s} \quad \text{and} \quad E_s(s_t^2) = \sigma^2.$$

Some readers may well ask: why should we not prefer  $s_t^2$  so defined in preference to  $s_z^2$  as defined by (vi) as our yard-stick of comparison when the end in view is to decide whether the column-means or row means yield an adequate estimate of variation in both dimensions of the grid? The fact is that we do—in a roundabout way—when we test the significance of a correlation ratio as in 16.08 below. Indeed, the issue is not referable to the arbitrament of common sense. In this context, a sufficient answer is that we shall require  $s_z^2$  for a different use, if the end in view is to assess the significance of differences between column- or row-means, when the universe is



not homogeneous in both dimensions ; but we may here anticipate another one. We shall later see that the ratio of  $s_c^2$  or of  $s_r^2$  to  $s_z^2$  is a Type VI distribution, whereas that of either to  $s_t^2$ , defined as above, is a Type I distribution. It happens that there are more convenient tables based on the former available for assessing the consistency of different estimates.

### 13.04 A BALANCE SHEET FOR TWO CRITERIA

As stated, we shall not at this stage examine the rationale of a test for the consistency of estimates of the variance of a putatively homogeneous universe. If the result of such a test, as explained in 16.05, leads us to reject the null hypothesis, we may then consider the question : what fractions of total variance in the universe of sampling are respectively attributable to agencies associated with the several criteria of classification and to residual sources ?

The credentials of any such balance sheet depend on a new set of assumptions, which we may specify under three headings :

- (a) *causal*, inasmuch as they refer to which effects of different components of variation contribute to a particular score value ;
- (b) *statistical*, inasmuch as they refer to the distribution of the score components singly or jointly ;
- (c) *operational*, inasmuch as they depend on the framework of repetition which the experimenter has in mind.

To clarify this threefold distinction, it will be helpful to cite a model experiment in which nature and nurture appear as the two criteria of classification. On six consecutive occasions with a 4-hour interval between any one and its predecessor or successor, a laboratory worker makes one determination of the blood calcium level of each of five rabbits, using the same five throughout. If we set out the 30 observations (scores) in a 5 (columns) by 6 (rows) table, we have to deal with three putative sources of variation :

- (i) a rhythm of variation within the 24-hour period in one and the same animal, its effect being therefore such as to increase variation of the row means ;
- (ii) systematic differences of the absolute level between animals at one and the same time, their effect being such as to increase variation of the column-means ;
- (iii) random errors of measurement sufficient to ensure cell to cell variation in the absence of either of the systematic components, and hence also some variation in row- and column-means.

It is admissible to conceive that each cell score in this set-up has three strictly additive components, which we shall refer to respectively as the residual, the column factor and the row factor. This constitutes a *causal* assumption. *Ex hypothesi*, the row factor varies from row to row, being constant from column to column and the column factor varies from column to column being constant from row to row within the sample ; but we are free to postulate random distribution of the residual from cell to cell in each dimension of the grid. This is a *statistical* assumption, as is the postulate that there is *zero covariance* between the residual and the other two components.\*

Neither the assumption of additivity nor that of zero covariance is necessarily true of any particular situation. They are attractive from a statistical viewpoint, because *the variance of the distribution of the sum of  $n$  variates is the sum of the variance of each, if the covariances are zero*. This circumstance makes it possible to express the total variance of a system as a sum of additive components ; and that indeed is what we mean when we speak of a balance sheet of variance.

\* The lay-out implies zero covariance of the row and column factors if their effects are strictly additive.

Having adopted these postulates with more or less plausibility we are not yet ready to proceed. For we have still to make an *operational* assumption, without which no unique solution is possible. Until we have decided within what framework of reference we choose to regard our experiment as a random sample, we are not in a position to undertake our analysis. In effect, this signifies that we have to find an answer to the question: in what way do we propose to repeat the experiment? One may repeat the experiment last cited in four ways:

- (i) by making  $n$  different determinations on each animal at one and the same time;
- (ii) by making observations at corresponding times in the course of the 24-hour period on the same set of rabbits on successive days;
- (iii) by making corresponding observations on more than one set of rabbits at identical times on one and the same day;
- (iv) by making corresponding observations on different sets of rabbits on successive days.

Evidently, the only use of an exhaustive balance sheet exhibiting components of variance is to prescribe what is likely to happen, if one does the same thing again. Evidently also, the sources of variation are not the same in the four ways which one might choose to regard as doing the same thing again in this context. The first implies that row and column factors remain constant throughout. The second and third respectively imply that the row factor alone or the column factor alone vary from one trial to another. The last signifies that both row and column factors vary. It leaves us free to postulate that they vary from sample to sample at random, and hence to invoke with more or less propriety a distribution law consonant with the possibility of assigning confidence limits to the entries of our balance sheet.

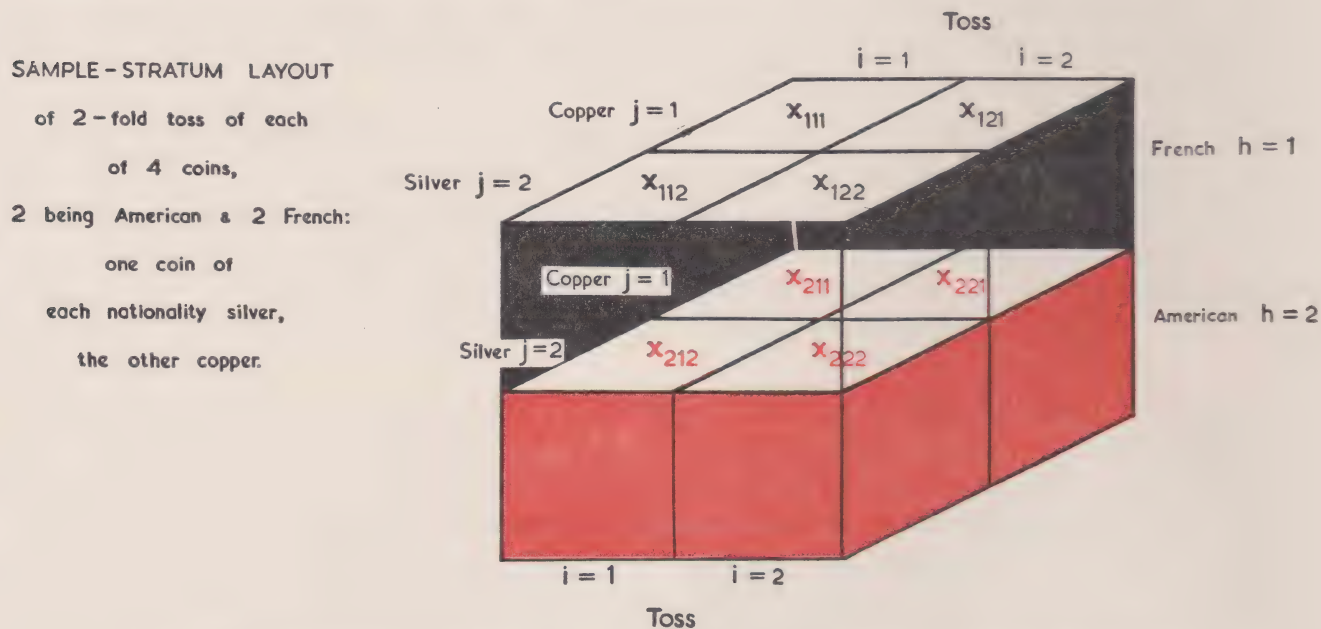


FIG. 102. Visualisation of a sample stratum for a lay-out involving 3 criteria of classification.

On this account, Churchill Eisenhart (*op. cit.*), who distinguishes between (i) and (iv) above as *Model I* and *Model II* situations, emphasises the distinction between them with particular reference to: (a) precautions taken to ensure random sampling in the design of the experiment; (b) whether the end in view is merely to assess the role of error variance or to effect a complete partition of the components of variation. Lee Crump (*op. cit.*), on the other hand, is more explicit about what we here regard to be the focal issue, *viz.* the operational intention.



The following assumptions are common to the treatment of the problem in accordance with the postulates of either Model I or Model II :

- (i) Three strictly additive components contribute to the sample cell-score  $x_{ij.s}$  in accordance with the following equation in which  $e_{ij.s}$  is the residual,  $F_{i.cs}$  the column factor and  $F_{j.rs}$  the row factor :

$$x_{ij.s} = e_{ij.s} + F_{i.cs} + F_{j.rs}.$$

- (ii) The covariance of any pair of components in (i.a) is zero, i.e.

$$\text{Cov}(e_{ij.s}, F_{i.cs}) = \text{Cov}(e_{ij.s}, F_{j.rs}) = \text{Cov}(F_{i.cs}, F_{j.rs}) = 0.$$

$x_{11.s} = e_{11.s}$	$x_{21.s} = e_{21.s}$	$x_{31.s} = e_{31.s}$
$x_{12.s} = e_{12.s}$	$x_{22.s} = e_{22.s}$	$x_{32.s} = e_{32.s}$
$x_{13.s} = e_{13.s}$	$x_{23.s} = e_{23.s}$	$x_{33.s} = e_{33.s}$

HOMOGENEOUS CASE

<u>Column Factor:-</u>		$F_{1,c}$	$F_{2,c}$	$F_{3,c}$
<u>Column(i):</u>		i = 1	i = 2	i = 3
<u>Row Factor</u>	<u>Row(j)</u>			
$F_{1,r}$	j = 1	$e_{11,s} + F_{1,c} + F_{1,r}$	$e_{21,s} + F_{2,c} + F_{1,r}$	$e_{31,s} + F_{3,c} + F_{1,r}$
$F_{2,r}$	j = 2	$e_{12,s} + F_{1,c} + F_{2,r}$	$e_{22,s} + F_{2,c} + F_{2,r}$	$e_{32,s} + F_{3,c} + F_{2,r}$
$F_{3,r}$	j = 3	$e_{13,s} + F_{1,c} + F_{3,r}$	$e_{23,s} + F_{2,c} + F_{3,r}$	$e_{33,s} + F_{3,c} + F_{3,r}$

MODEL I

Column Factor  $F_{i.c}$  constant within column of sample (layer) and within column-slab of universe (3-dimensional grid)

Row Factor  $F_{j.r}$  constant within row of sample (layer) and within row-slab of universe (3-dimensional grid)

Column Factor:	$F_{1.cs}$	$F_{2.cs}$	$F_{3.cs}$	
Column(i):	i = 1	i = 2	i = 3	
Row Factor	Row(j)			
$F_{1.rs}$	j = 1	$e_{11.s} + F_{1.cs} + F_{1.rs}$	$e_{21.s} + F_{2.cs} + F_{1.rs}$	$e_{31.s} + F_{3.cs} + F_{1.rs}$
$F_{2.rs}$	j = 2	$e_{12.s} + F_{1.cs} + F_{2.rs}$	$e_{22.s} + F_{2.cs} + F_{2.rs}$	$e_{32.s} + F_{3.cs} + F_{2.rs}$
$F_{3.rs}$	j = 3	$e_{13.s} + F_{1.cs} + F_{3.rs}$	$e_{23.s} + F_{2.cs} + F_{3.rs}$	$e_{33.s} + F_{3.cs} + F_{3.rs}$

MODEL II

Column Factor  $F_{i.cs}$  constant within column of sample (layer) variable within column-slab of universe (3-dimensional grid) row-slab and column-slab distributions identical with one another and with that of whole grid

Row Factor  $F_{j.rs}$  constant within row of sample (layer) variable within row-slab of universe (3-dimensional grid) column-slab and row-slab distributions identical with one another and with that of whole grid

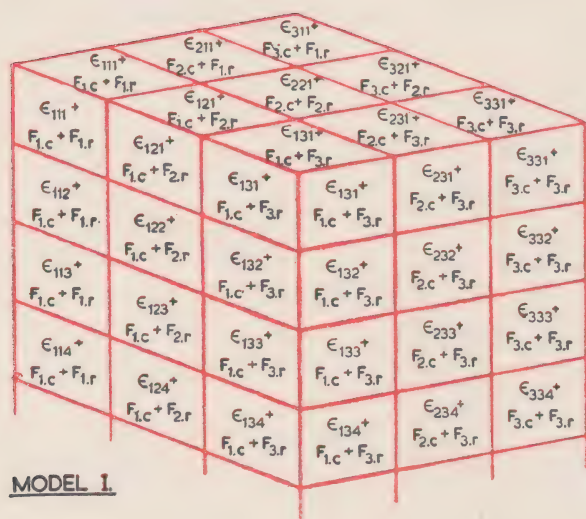
FIG. 103. Two sets of assumptions concerning additive score components.

- (iii) If  $\sigma^2 = V_x$  is the variance of the distribution of the total score  $x_{ijk}$  in the universe, and  $\sigma_e^2$ ,  $\sigma_c^2$ ,  $\sigma_r^2$  are the corresponding variances of the distributions of the score components in (i) it follows that

$$\sigma^2 = \sigma_e^2 + \sigma_c^2 + \sigma_r^2.$$

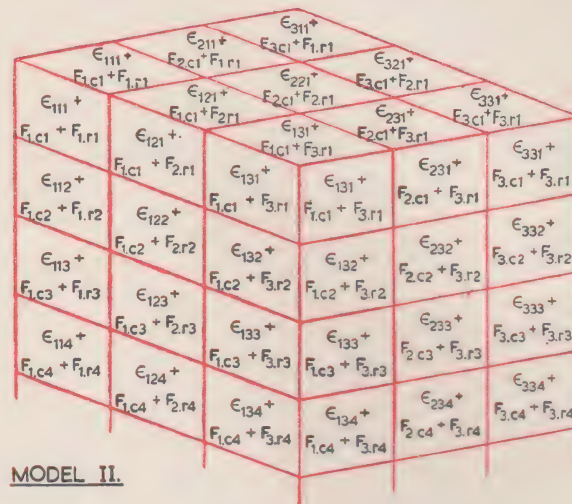
- (iv) The residual component  $e_{ijk}$  varies from cell to cell within the row and within the column of the sample random-wise, so that the distribution of residual score components is the same in all pillars of the 3-dimensional grid of the complete random sample distribution.
- (v) Within the same layer of the 3-dimensional grid  $F_{i.cs}$  varies from cell to cell *only* within the row, being fixed for the column, and  $F_{j.rs}$  varies from cell to cell *only* within the column, being fixed within the row.
- (vi) Accordingly, the distribution of the column factor in the sample as a whole is identical with its row distribution, all rows being alike with reference thereto; and the

**3 × 3 CLASS UNIVERSE STRATIFIED IN 2 DIMENSIONS**  
by addition of fixed Row and Column increments within the layer



Column Factor  $F_{1,c}$  constant within column of sample (layer)  
and within column-slab of universe (3-dimensional grid)

Row Factor  $F_{1,r}$  constant within row of sample (layer)  
and within row-slab of universe (3-dimensional grid)



Column Factor  $F_{1,c}$  constant within column of sample (layer)  
variable within column-slab of universe (3-dimensional grid)  
row-slab and column-slab distributions identical with one  
another and with that of whole grid

Row Factor  $F_{j,r}$  constant within row of sample (layer)  
variable within row-slab of universe (3-dimensional grid)  
column-slab and row-slab distributions identical with  
one another and with that of whole grid

FIG. 104. The 3-dimensional universe of sampling for the two model situations of Fig. 103.

distribution of the row factor in the sample as a whole is identical with its column distribution, all columns being in this respect alike.

The postulates peculiar to Model I are

- (vii)  $F_{i,cs}$  is fixed for all cells within the same column-slab as well as for all cells within the same column of the same layer, and  $F_{j,rs}$  is fixed for all cells within the same row-slab as well as for all cells of the same row within the same layer.
- (viii) Hence the variance of the row factors within a column as within a layer is  $\sigma_r^2$  and the variance of the column factors within a row as within a layer is  $\sigma_c^2$ .

Contrariwise, the postulates of Model II will be that  $F_{i,cs}$  and  $F_{j,rs}$  vary at random from layer to layer in the sense that

- (ix) each row-slab and each pillar therein accommodates a complete random distribution of column factors identical with the distribution of column factors in the 3-dimensional grid as a whole, whence also in virtue of (v) and (vi) identical with the distribution of column factors in the column-slab;
- (x) each column-slab and each pillar therein accommodates a complete random distribution of row factors identical with the distribution of column factors in the whole 3-dimensional grid, whence likewise in virtue of (v) and (vi) identical with the row-slab distribution of row factors.

In what follows we shall first explore the consequences of Model II. The only new consideration of moment arising from the foregoing definitions is that the whole  $rc$ -fold sample of  $x$ -scores which supplies us with an  $rc$ -fold sample of  $e$ -scores is a  $c$ -fold sample of row factors



on account of the identity of the rows with respect to the latter and an  $r$ -fold sample of column factors on account of the identity of the columns with respect thereto. We may express this otherwise by saying that the sample as a whole furnishes us with no information about the column factors other than what we may infer from the composition of any one of the rows alone, and no information about the row factors other than what we may infer from any one of the columns equally.

In accordance with Model II postulates, we shall need symbols for the variance of the score components as below :

	<i>Residual</i>	<i>Row factor</i>	<i>Column factor</i>
<i>Whole sample</i> (layer) . . .	$V_{e.s}$	$V_{r.s} = V_{r.cs}$	$V_{c.s} = V_{c.rs}$
<i>Within-row</i> . . . . .	$V_{e.rs}$	$V_{r.rs} = 0$	$V_{c.rs}$
<i>Within-column</i> . . . . .	$V_{e.cs}$	$V_{r.cs}$	$V_{c.cs} = 0$

From what has been said, the expected values of the above are

$$\begin{aligned}
 E_s(V_{e.s}) &= \frac{rc-1}{rc} \sigma_e^2 & E_s(V_{r.s}) &= \frac{r-1}{r} \sigma_r^2 & E_s(V_{c.s}) &= \frac{c-1}{c} \sigma_c^2; \\
 E_s(V_{e.rs}) &= \frac{c-1}{c} \sigma_e^2 & 0 & & E_s(V_{c.rs}) &= \frac{c-1}{c} \sigma_c^2; \\
 E_s(V_{e.cs}) &= \frac{r-1}{r} \sigma_e^2 & E_s(V_{r.cs}) &= \frac{r-1}{r} \sigma_r^2 & 0. &
 \end{aligned}$$

We now recall the procedure of which the following is a pattern :

$$E_s.M(V_{e.cs}) = E_s.E_c(V_{e.cs}) = E_c.E_s(V_{e.cs}) = \frac{r-1}{r} \sigma_e^2.$$

If the components have zero covariance

$$\begin{aligned}
 V_{x.s} &= V_{e.s} + V_{c.s} + V_{r.s}; \\
 V_{x.rs} &= V_{e.rs} + V_{c.rs}; \\
 V_{x.cs} &= V_{e.cs} + V_{r.cs}.
 \end{aligned}$$

Whence we derive :

$$\begin{aligned}
 E_s(V_{x.s}) &= \frac{rc-1}{rc} \sigma_e^2 + \frac{r-1}{r} \sigma_r^2 + \frac{c-1}{c} \sigma_c^2; \\
 E_s.M(V_{x.rs}) &= \frac{c-1}{c} \sigma_e^2 + \frac{c-1}{c} \sigma_c^2; \\
 E_s.M(V_{x.cs}) &= \frac{r-1}{r} \sigma_e^2 + \frac{r-1}{r} \sigma_r^2.
 \end{aligned}$$

If  $V_z$  has the same meaning as in 13.03 :

$$\begin{aligned}
 E_s(V_z) &= E_s.M(V_{x.rs}) + E_s.M(V_{x.cs}) - E_s(V_{x.s}) \quad . \quad . \quad . \quad (i) \\
 &= \left( \frac{c-1}{c} + \frac{r-1}{r} - \frac{rc-1}{rc} \right) \sigma_e^2; \\
 E_s(V_z) &= \frac{(r-1)(c-1)}{rc} \sigma_e^2.
 \end{aligned}$$

<i>Component</i>	<i>Unbiased Estimate</i>
$\sigma_c^2$	$\frac{S_c - S}{(r-1)(c-1)} + \frac{S_r - S_q}{r(r-1)(c-1)};$
$\sigma_r^2$	$\frac{S_r - S}{(r-1)(c-1)} + \frac{S_c - S_q}{c(c-1)(r-1)};$
$\sigma^2$	$\frac{S + S_q - S_c - S_r}{(r-1)(c-1)}.$





## 13.05 THE ADDITIVE PRINCIPLE

In seeking a rationale for the construction of a balance sheet of variation we have postulated a universe of scores with 3 additive components, a column factor, a row factor and a residual. On the assumption that there is zero covariance between any pair of them, the true variance of the composite score in the universe of choice is the sum of the variances of the three components, i.e.  $\sigma^2 = \sigma_e^2 + \sigma_r^2 + \sigma_c^2$ . Thus  $\sigma_e^2$  stands for what the total variance would be if there were no source of systematic variation associated with the row and column criteria of classification. The tidiness of this relation has a deceptive air of finality. So it is important that the student should understand what factual assumptions entitle us to construct a balance sheet in accordance with the algebraic postulates of 13.04.

From the factual viewpoint, the pivotal postulate is that the effects of both row and column factors are *strictly additive*. This signifies that effects of sources of variation associated with the two-class systems are such as to change the mean value of the row or column score distribution without changing its form or scale. Now it is easy to imagine many other ways in which inter-class variation might arise. In much experimental work, change of scale or dispersion without change of mean in virtue of the competence of the worker is just as likely an assumption, perhaps more so. Hence the attractiveness of the additive postulate resides less in its relevance to external nature than in its convenience to the mathematician. Of this, as of other assumptions commonly made in relation to the same class of procedure, we may cite the comment of Churchill Eisenhart (1947): "the only motivation that has been given is the more general nature of the inferences that may be drawn . . . when it is satisfied".

In any real situation, it therefore behoves us to ask whether the additive postulate is indeed plausible; and it is conspicuously open to question in the field of earliest and most extensive applications of variance analysis. Here again the remarks of Churchill Eisenhart (*op. cit.*) are explicit and salutary:

Hence, when additivity does not prevail we say that there are interactions between row and column factors. Thus, in the case of varieties and treatments . . . additivity implies that, under the general experimental conditions of the test, the true mean yield of one variety is greater (or less) than the true mean yield of another variety by an amount—an additive constant, not a multiplier—that is the same for each of the treatments concerned, and, conversely, the true mean yield with one treatment is greater (or less) than the true mean yield with another treatment by an amount which does not depend upon the variety concerned; which is exactly what is meant when we say that there are no "interactions" between varietal and treatment effects.

Mathematicians who are not conversant with the vagaries of gene exhibition and biologists who are not at home with the technical intricacies of the thesis expounded by the writer of the remarks cited above will not regard it as unprofitable to pinpoint what is of cardinal importance to the present discussion by reference to a naturalistic illustration. Our supposition is that we record in 3 different environments the size (*yield*) attained at a given age by individuals of two species of flowering plants, one (*A*) being calciphil and the other (*B*) being calciphobe. The three environments (*treatments*) are then the native soil (untreated), native soil with addition of a neutral calcium salt and native soil treated with a neutral potassium salt. To drive home the point Churchill Eisenhart makes in his reference to treatment (*nurture*), variety (*nature*) and yield (*phenotype*), we may disregard the residual component of the cell-score (*yield*) arising from random errors of measurement and uncontrolled subsidiary difference w.r.t. environment. If we denote the cell-score in the absence of residual error so defined as  $u_{ij}$ , the column (*species*) factors respectively by  $F_a$  and  $F_b$ , and the row (treatment) factors as  $F_1$ ,  $F_2$  and  $F_3$ , our set-up as prescribed by the additive postulate is



	Species <i>A</i>	Species <i>B</i>
Untreated ( <i>r</i> = 1)	$u_{11} = (F_a + F_1)$	$u_{21} = (F_b + F_1)$
Excess Ca ( <i>r</i> = 2)	$u_{12} = (F_a + F_2)$	$u_{22} = (F_b + F_2)$
Excess K ( <i>r</i> = 3)	$u_{13} = (F_a + F_3)$	$u_{23} = (F_b + F_3)$

The implications of this become more obvious, if we set the result out thus :

	Species <i>A</i>	Species <i>B</i>
Effect of Ca . . .	$(u_{12} - u_{11}) = (F_2 - F_1)$	$= (u_{22} - u_{21})$
Effect of K . . .	$(u_{13} - u_{11}) = (F_3 - F_1)$	$= (u_{23} - u_{21})$

The above schema signifies that a fixed excess of *Ca* increases the size of *B* and *A* by an equal amount, a statement which is inconsistent with our own initial assumption that the two species are respectively calciphobe and calciphil. Likewise the additive postulate signifies in this context that *B* and *A* react by equal responses to a fixed increment of *K*, an assertion inconsistent with general experience of ionic antagonisms in the biological domain. On the contrary, we should expect a calciphil species, which reacts by increased yield to increase of *Ca* soil content, to react by diminished yield to excess of *K*, and a calciphobe species which reacts by diminished yield to increase of the *Ca* soil content, to react by increased yield to excess of *K*.

From the field of interspecific variation, it would be possible to cite numerous examples of comparable situations, but the writer has sufficiently emphasised their occurrence within the domain of intraspecific variation. More recently Haldane\* has classified known types of interaction, i.e. departure from the assumption of additivity, by recourse to experimental data. It is indeed open to question whether there exists any nature-nurture situation about which we can make any such assumptions with confidence in the absence of corroboration, or whether it will often happen that such an assumption is plausible. What is certain, as illustrated by the foregoing example, is likewise embodied in the adage that one man's meat is another man's poison. Many situations arise in which stimulus *X* increases response of genotype *A* and inhibits that of genotype *B*, while stimulus *Y* decreases response of genotype *A* and augments that of genotype *B*. That it is necessary to remind the biologist familiar with his materials that the additive postulate may therefore be grossly inapplicable to a set-up in which the two criteria classification are nature and nurture, is because relatively few biologists who invoke the technique under discussion with a view to the construction of a balance sheet exhibiting the respective contributions of nature and nurture variables realise that the additive postulate is in fact the keystone of the entire edifice.

Accordingly, we may thus sum up the outcome of our enquiry at this stage :

- (1) The possibility of constructing a true bill which sets out what fractions of total population variance are respectively attributable to one or other source of variation specified by the class criteria and to residual errors of observation or other uncontrolled circumstances presupposes the truth of the postulate that the components are additive ;
- (2) From inspection of the data of a single small-scale experiment it is never possible to infer with certainty that this postulate is valid, and there will arise many situations in which it is grossly incorrect.

\* J. B. S. Haldane (1946), "The Interaction of Nature and Nurture". *Ann. Eugen.* 13, 197.



These considerations prompt us to ask: is it possible to justify the additive postulate by recourse to observation, and if so, how? To answer this, we shall suppose that the joint contribution of  $F_{i.cs}$  and  $F_{j.rs}$  to the score value exceeds or falls short of their sum by an amount  $F_{ij.s}$  which varies from cell to cell, i.e.:

$$x_{ij.s} = e_{ij.s} + F_{i.cs} + F_{j.rs} + F_{ij.s}.$$

Evidently the expected value of  $s_z^2$  in 13.04 will not be  $\sigma_e^2$  unless  $F_{ij.s} = 0$ , and the design of an experiment involving single score values for each cell in a 2-way lay-out provides no occasion for distinguishing between two components which both vary from cell to cell. On the other hand, their effects are distinguishable if we resort to *n*-fold replication, i.e. *n*-fold repetition of each observation without changing the row and column sources of variation. In such an experimental design, we conceive our sample as a stratum of *n* layers. The residuals vary random-wise from cell to cell within a layer and from layer to layer within a pillar. If the replication is faithful, the component  $F_{ij.s}$  varies from cell to cell within a layer but not within a pillar. Accordingly, we can ask whether the measures of cell-to-cell variation within a pillar and within a stratum are consistent, i.e. if  $F_{ij.s} = 0$ , in a set-up for 3 criteria of classification involving replication as the new one.

We shall examine this issue in 13.06. Here it suffices to point out that the construction of a balance sheet for an experiment involving replication is valid only if: (a) the analysis fails to disclose a new component of variation; (b) we have other reasons for assuming that the 3 systematic components conform to the postulates of additivity and zero covariance. If the results of identical replication do *not* confirm the assumption that the postulate is valid, the inclusion of a separate component of *interaction* as defined by Churchill Eisenhart in the balance sheet of causality merely serves to announce that the procedure for constructing it is defective, hence also that it is not a true bill.

Before we examine the credentials of the balance sheet for a replication experiment, it is fitting to examine what we may rightly infer, if there is indeed good reason to believe that the additive principle holds good. We may then interpret our balance sheet as

- (i) a recipe for assigning to what errors mean measurements are subject when we exclude one or other source of variation;
- (ii) an overall picture of how much variability remains when we eliminate one or other source.

To clarify the meaning of (i), the illustrative experiment already cited will serve our purpose. At a given time of day, the data supply us with a mean figure for the blood calcium level of different rabbits. This figure is therefore subject both to residual sampling error inherent in the method of measurement and to variation arising from the fact that different measurements refer to 5 different individuals. The unbiased estimate of the residual variance being  $s_e^2$ , that of the mean of a 5-fold sample is  $\frac{1}{5}(s_e^2)$  in virtue of (vii) in 7.03 of Vol. I. Alternatively, we may ask what would be the sampling variance for the mean of the series of 6 determinations on the same rabbit at different times of day or night, i.e. to what sampling variance our column means referable to the same rabbit are subject as the result of errors of measurement alone. In this case, our concern is with the mean of a 6-fold sample, and the required parameter is  $\frac{1}{6}(s_e^2)$ . In general, we may say that the mean row-scores and the mean column-scores are respectively subject to sampling variance of  $(s_e^2 \div c)$  and  $(s_e^2 \div r)$ . In the writer's view, this is the most fruitful use of the procedures subsumed by the term analysis of variance, if only because it operates within the domain of *estimation* and therefore entails none of the debatable issues raised by recent controversy concerning decisions made within the framework of a unique null hypothesis.



An alternative conception of the sort of questions an accredited Balance Sheet of Variance may answer brings into focus an important difference between *Model I* and *Model II* of 13.04. As an assemblage of unbiased estimates of universe components of variance, the balance sheet would appear to be just as valid, if constructed on one or other assumption ; but we may wish to take the further step of placing confidence limits (*see* 16.03 below) around each of our estimates of the components. To do so, we must then invoke certain assumptions concerning the distribution of the row and column factors. If we scrutinise our experimental data through the spectacles of *Model II*, we are free to postulate with more or less justification a normal distribution of all three-score components in the universe as a whole. Thereafter the problem stated is purely mathematical, if we are entitled to regard the choice of sample as random.

American writers on analysis of variance are not slow to stress the fact that random choice of column- or row-score components is often inconsistent with experimental design, as in the following remarks of Churchill Eisenhart :

“ . . . when an experimenter selects two or more treatments, or two or more varieties, for testing, he rarely, if ever, draws them at random from a population of possible treatments or varieties ; he selects those that he believes are most promising. Accordingly *Model I* is generally appropriate where treatment, or variety comparisons are involved. On the other hand, when an experimenter selects a sample of animals from a herd or a species, for a study of the effects of various treatments, he can insure that they are a random sample from the herd, by introducing randomization into the sampling procedure, for example, by using a table of random numbers. But he may consider such a sample to be a random sample from the species, only by making the assumption that the herd itself is a random sample from the species. In such a case, if several herds (from the same species) are involved, *Model II* would clearly be appropriate with respect to the variation among the animals from each of the respective herds, and might be appropriate with respect to the variation of the herds from one another.”

Lee Crump (1946) writes in the same vein :

“ *A Note of Warning.* It must be remembered that in using the analysis of variance to estimate variance components, we have assumed the elements of the fundamental equation to be randomly selected from an infinite population. In an experiment where three widths of spacing some crop are purposely selected for trial, it is not reasonable to regard these widths as random samples from all possible widths. On the other hand the blocks in a field experiment may sometimes quite reasonably be regarded as a random sample of all such blocks. In sampling production from, say, three machines in a factory, where these machines constitute all the machines which the factory has or is likely to have, it is more reasonable to regard these machines as the whole of a finite population than to consider them as random samples from some infinite population. If the factory owner is sampling production with a view to purchasing more machines of the same type, the three machines may be appropriately regarded as samples of the infinite population made up of all machines of the same type.”

The more reasonable attitude to the three machines as the whole of a finite population is in fact a *Model I* view of the situation ; but the same example also brings into focus a semantic difficulty which besets justification of the alternative view. Indeed, the foregoing remarks of Churchill Eisenhart resolve only part of the difficulty of justifying the assumption that choice of classificatory variables is truly random. To be sure, we can choose cows of a particular herd at random, but if we do, our assessment legitimately refers only to that herd. To extend it justifiably to others of the same breed, we have to invoke the additional assumption that the range of intra-specific variation does not materially differ from herd to herd ; and to say that this assumption is itself unjustifiable deprives the assessment of public utility. In the nature of things, random choice of fertilisers of all possible chemicals curiosity or perversity may prompt the investigator



to add to the soil is a concept devoid of operational meaning ; and random choice of varieties within a species or of individuals within a variety is a concept we can justify without recourse to a God's eye view of the universe only as a description of a local set-up. For a truly random choice of rabbit varieties in Kent would not be a truly random choice of rabbits in Kentucky.

T. H. Huxley remarked rightly that mathematics is a mill which cannot grind out ingredients other than those put into it. What is true of any statistical technique is therefore true of analysis of variance, and especially so. For statisticians of an earlier vintage never identified their terms of reference with so ambitious a title as the *design of experiments*. No statistical technique is an adequate substitute for common sense, alertness to the nature of the problem on the part of those who ought to be clear about it or for ingenuity directed to the removal of irrelevant variables in an experimental set-up. Indeed, it is well to remind ourselves that experimental science, in its assault on problems most successfully attacked by experimental methods to date, had advanced far towards its present stature without recourse to statistical principles of any sort.

It is therefore necessary to insist that analysis of variance—like any other sort of statistical procedure—has a limited sphere of usefulness, especially because its legitimate uses are at present difficult to assess against a background of novel logical premises and, for most of us, unfamiliar mathematical procedures. In the situation we have used as an illustration of a 2-way classificatory set-up, our assumption has been that the investigator wishes to ascertain with as great economy as possible whether the blood calcium level of rabbits is or is not subject to a diurnal rhythm, i.e. a rise and fall within a 24-hour period. A balance sheet of variance which exhibits a significantly large component w.r.t. observations on different animals at different times of the day does not in fact answer the question last stated, such a result being consistent with a quite erratic fluctuation during a particular time interval such as 24 hours. In so far as the analysis bears on the problem stated, it is helpful because we can state to how much sampling variance our mean values for determination at different times of day are subject when we have eliminated all sources of individual variation. Hence we can see whether there is a consistent trend of our mean values throughout a 24-hour period without recourse to the more homely custom of repeated experiments of the same sort. In such a situation, the experimentalist is entitled to prefer the assurance of a consistent answer by recourse to laboratory experience of several days duration to the consolations of mathematics ; but there may well arise situations in the practice of industry or in sociological enquiry such as to commend a *first approach* which is more economical.

With full recognition of the existence of situations in which an economical preview is indeed advantageous, it remains none the less true that no statistical procedure can rightly claim to provide a rationale for the design of experiments regardless of the end in view ; and a widely quoted illustration of the use of analysis of variance in particular is instructive as a warning against any such mechanical view of the value of statistical methods. In an early issue of *Biometrika*, Oswald Latter (1902) published the result of measuring the length of 1572 eggs of the cuckoo including 264 assignable to known foster parents of 6 different species. The odds are in fact about 100 : 1 that variation between nests of one or other type is not wholly explicable in terms of variation within nests. Since 1902 the same set of figures has passed from one textbook to another to illustrate one or other statistical technique fashionable at the time, latterly as an illustration of homogeneity tests w.r.t. a one-way classification involving unequal sample numbers as in 13.07. Indeed, the writer of a comparatively recent book on statistics for sociologists introduces the topic with the complaint that “ it is a considerable jump from lengths of eggs in a cuckoo's nest . . . to sociological problems ”.\*

\* Margaret Jarman Hagood : *Statistics for Sociologists* (1941).



The slip-up in the last sentence cited is pardonable, since sociologists are under no obligation to know that cuckoos do not have nests, unless current texts which trade in this exhibit disclose what is the end in view. On the other hand, it would be difficult to offer the naturalist an example of the use of statistics less calculated to inspire confidence. Similarities with respect both to colour and form between the cuckoo's egg and those of the foster parent had long been a matter of comment and discussion among enthusiastic bird watchers and egg collectors. It was also well known that cuckoo eggs have a character peculiar to the locality where found, as discussed by Eugene Rey (1892) in his book *Old and New information concerning the domestic economy of the Cuckoo*. One may presume that Latter, himself a first-rate naturalist, knew all this when he chose the topic; and one may be confident that he would have been able to throw further light on it if the hypnotic influence of Pearson's apotheosis of measurement as an end in itself had not enlisted his industry in an undertaking unlikely to add anything to what was already commonplace among bird watchers.

Indeed, it is scarcely too much to say that no author who uses Latter's data as exemplary material has been able to convince the reader that the outcome of the ensuing arithmetical exploits has greatly advanced biological knowledge. It is also safe to say that it put at the disposal of those who have later clarified the enigma no helpful clue for their use or guidance. The facts, disclosed by E. P. Chance (*The Truth about the Cuckoo*, 1940) as we now know them, are the result of painstaking observations on the behaviour of individual cuckoos during the same and successive seasons. Briefly they are as follows. The same cuckoo returns in successive years to the same territory and almost invariably lays its eggs in the nest of a particular species. All the available evidence points to the conclusion that a cuckoo reared in a particular *territory* mates with another cuckoo reared in the same territory. In short, cuckoos are divisible into local sub-species each with its dominant foster-parent type, and selection has presumably ensured the survival of genotypes most fitted to lay eggs acceptable to the latter.

Thus the truth about the cuckoo as it here concerns us is that a much-publicised statistical enquiry did not in fact draw attention to a new problem, and it did little if anything to clarify one which field naturalists already recognised as such. It is not easy to see how it would have been possible to elicit the relevant facts by methods other than *intensive* work of field observers, for the most part allergic to statistics of any sort. Statisticians who wish to enlist greater respect for the proper uses of statistics would therefore be wise to refrain from further comment on the cuckoo question when their aim is to show how statistics can help the field worker.

One *caveat* it is still necessary to emphasise in this context concerns a class of judgments common to many situations involving *multiple*, as opposed to binary, classification. So long as our preoccupation is with only two classes the issue of homogeneity is straightforward. Either the statistical data referable to the two samples are indicative of a real difference or they are not. When we turn our attention to a system of more than 2 classes the assertion that there exists a real inter-class difference may signify at opposite extremes: (i) a graded effect distinguishing any one class from every other; (ii) a clear-cut threshold response which may differentiate only one class from any other. We meet with clear-cut *threshold* effects very commonly in biological enquiry; and we have no reason to disregard the possibility of doing so in social science. Where this is so, a multiple classification of the data may conceal or obscure a real difference which two-fold division at an appropriate level would bring sharply into focus. In the last resort, any statistical technique referable to a system of many classes will be more or less useful to the extent that the investigator exercises good judgment of his materials in the initial task of classifying them.



## 13.06 BALANCE SHEET FOR THREE CRITERIA

An analysis involving three criteria of classification may be *replicative* or *complete*. The first, sometimes referred to as *incomplete* 3-factor analysis, signifies that the third criterion of classification is simply *repetition*, as when we have  $n$  score values for each of  $rc$  cells in a lay-out involving two specific taxonomical categories. The second, referred to as complete three-factor analysis (without interaction), signifies that each of the  $n$  observations constitutes a member of a class, as in the coin model of Fig. 102.

To clarify the *replicative* case we may imagine an experimental design of the following type. The scores  $a_{ij}$  and  $b_{ij}$  respectively refer to the red cell count of different blood samples from one and the same female rabbit at one and the same time of day :

	Rabbit I	Rabbit II
12 noon	$a_{11} ; b_{11}$	$a_{21} ; b_{21}$
12 midnight	$a_{12} ; b_{12}$	$a_{22} ; b_{22}$

In this set-up we have initially 2 specific criteria of classification: type of individual (*columns*) and time of day (*rows*). Precise repetition should lead to consistent estimates both of the error variance and that of the putatively additive row and column factors, if there is indeed zero covariance between the row and column factors, though a cell factor indistinguishable from the residual variance in a single trial would be separable in a repetition *involving no new source of systematic variation*. In that event, we should be able to distinguish from true error, which varies from cell to cell in any one experiment and from cell to corresponding cell in successive experiments, a component which varies only from cell to cell in any single experiment being constant in corresponding cells of successive experiments. The words in italics are the operative phrase in the sentence above. In addition to residual sources of variation arising from errors in connexion with each of the 8 counts involved, we may conceive a *systematic* source of error introduced by defective procedure, e.g. the use of a separate syringe needle for each rabbit at each time of day. We may then speak of a cell factor  $F_{ij.s}$  which varies from cell to cell like the residual  $e_{ij.s}$  of 13.04 but is constant within the cell. To represent this conception we need to label a third (within-cell) dimension ( $h = 1, 2 \dots n$ ) of variation and our score components as follows :

$$\begin{array}{ll}
 \text{Column Factor} & \cdot \quad F_{i.cs} \\
 \text{Cell Factor} & \cdot \quad \cdot \quad F_{ij.s}
 \end{array}
 \qquad
 \begin{array}{ll}
 \text{Row Factor} & \cdot \quad \cdot \quad F_{j.rs} \\
 \text{Residual} & \cdot \quad \cdot \quad e_{hij.s}
 \end{array}$$

We may then visualise the foregoing lay-out as below :

$a_{11} = e_{111.s} + F_{11.s} + F_{1.cs} + F_{1.rs}$	$a_{21} = e_{121.s} + F_{21.s} + F_{2.cs} + F_{1.rs}$
$b_{11} = e_{211.s} + F_{11.s} + F_{1.cs} + F_{1.rs}$	$b_{21} = e_{221.s} + F_{21.s} + F_{2.cs} + F_{1.rs}$
$a_{12} = e_{112.s} + F_{12.s} + F_{1.cs} + F_{2.rs}$	$a_{22} = e_{122.s} + F_{22.s} + F_{2.cs} + F_{2.rs}$
$b_{12} = e_{212.s} + F_{12.s} + F_{1.cs} + F_{2.rs}$	$b_{22} = e_{222.s} + F_{22.s} + F_{2.cs} + F_{2.rs}$

From a formal point of view, the outcome of the analysis will be the same whether we interpret  $F_{ij.s}$  as: (a) an unsuspected independent systematic source of variation,









We now recall the statistic  $V_{zn}$  defined by (xxv) of 11.05, viz.:

$$V_{zn} = V_{x.s} - V(M_{x.cs}) - V(M_{x.rs}) - M(V_{x.rcs}).$$

Whence from (iii)-(v) :

$$E_s(V_{zn}) = \frac{(r-1)(c-1)}{ncr} \sigma_e^2 + \frac{(r-1)(c-1)}{cr} \sigma_{cr}^2 \quad . \quad . \quad . \quad (vi)$$

Accordingly, we may define the following statistics :

$$\begin{aligned} E_s(s_c^2) &= \sigma_e^2 + n\sigma_{cr}^2 + nr\sigma_c^2 & \text{if } s_c^2 &= ncr \cdot V(M_{x.cs}) \div (c-1); \\ E_s(s_r^2) &= \sigma_e^2 + n\sigma_{cr}^2 + nc\sigma_r^2 & \text{if } s_r^2 &= ncr \cdot V(M_{x.rs}) \div (r-1); \\ E_s(s_{zn}^2) &= \sigma_e^2 + n\sigma_{cr}^2 & \text{if } s_{zn}^2 &= ncr \cdot V_{zn} \div (r-1)(c-1); \\ E_s(s_{cr}^2) &= \sigma_e^2 & \text{if } s_{cr}^2 &= ncr \cdot M(V_{x.rcs}) \div cr(n-1). \end{aligned}$$

For purposes of computation, we may set out the foregoing results in accordance with the schema of (xxv)-(xxxi) in 11.05 :

Mean Sums of Squares	Divisor	Estimate	Expected Value
$S_e - S$	$c - 1$	$s_c^2$	$\sigma_e^2 + n\sigma_{cr}^2 + nr\sigma_c^2$
$S_r - S$	$r - 1$	$s_r^2$	$\sigma_e^2 + n\sigma_{cr}^2 + nc\sigma_r^2$
$S + S_{cr} - S_e - S_r$	$(r-1)(c-1)$	$s_{zn}^2$	$\sigma_e^2 + n\sigma_{cr}^2$
$S_q - S_{cr}$	$rc(n-1)$	$s_{cr}^2$	$\sigma_e^2$

### Complete 3-factor Analysis without interaction:

If we proceed in accordance with (ii) above, we have

$$\begin{aligned} E_s \cdot V(M_{x.cs}) &= E_s(V_{x.s}) - E_s M(V_{x.cs}) \\ &= \frac{ncr-1}{ncr} \sigma_e^2 + \frac{c-1}{c} \sigma_e^2 + \frac{r-1}{r} \sigma_r^2 + \frac{n-1}{n} \sigma_n^2 \\ &\quad - \frac{nr-1}{nr} \sigma_e^2 - \frac{r-1}{r} \sigma_r^2 - \frac{n-1}{n} \sigma_n^2, \\ \therefore E_s \cdot V(M_{x.cs}) &= \frac{c-1}{ncr} \sigma_e^2 + \frac{c-1}{c} \sigma_e^2 \quad . \quad . \quad . \quad . \quad . \quad . \quad (vii) \end{aligned}$$

In the same way we derive

$$E_s \cdot V(M_{x.rs}) = \frac{r-1}{ncr} \sigma_e^2 + \frac{r-1}{r} \sigma_r^2 \quad . \quad . \quad . \quad . \quad (viii)$$

$$E_s \cdot V(M_{x.ns}) = \frac{n-1}{ncr} \sigma_e^2 + \frac{n-1}{n} \sigma_n^2 \quad . \quad . \quad . \quad . \quad (ix)$$

Hitherto, we have based one item of our balance sheet on the difference between the total variance and the parameters used to evaluate all the remaining estimates. So we now define

$$V_L = V_{x.s} - V(M_{x.cs}) - V(M_{x.rs}) - V(M_{x.ns}).$$





For we can postulate 3 systematic cell effects in addition and test the hypothesis that each is negligible. For simplicity let us assume that we have other grounds for believing that there is no interaction between  $F_{h.ns}$  and either  $F_{i.cs}$  or  $F_{j.rs}$ . Our concern is then with the possibility of interaction between  $F_{i.ns}$  and  $F_{j.rs}$ . Accordingly we postulate :

$$x_{hij.s} = e_{hij.s} + F_{i.cs} + F_{j.rs} + F_{h.ns} + F_{ij.s}.$$

By the foregoing procedure we should then derive

$$E_s . V_{x.s} = \frac{ncr - 1}{ncr} \sigma_e^2 + \frac{c - 1}{c} \sigma_c^2 + \frac{r - 1}{r} \sigma_r^2 + \frac{n - 1}{n} \sigma_n^2 + \frac{cr - 1}{cr} \sigma_{cr}^2;$$

$$E_s . V(M_{x.cs}) = \frac{c - 1}{ncr} \sigma_e^2 + \frac{c - 1}{cr} \sigma_{cr}^2 + \frac{c - 1}{c} \sigma_c^2;$$

$$E_s . V(M_{x.rs}) = \frac{r - 1}{ncr} \sigma_e^2 + \frac{r - 1}{cr} \sigma_{cr}^2 + \frac{r - 1}{r} \sigma_r^2;$$

$$E_s . V(M_{x.ns}) = \frac{n - 1}{ncr} \sigma_e^2 + \frac{n - 1}{n} \sigma_n^2;$$

$$E_s . V_{zn} = \frac{(r - 1)(c - 1)}{ncr} \sigma_e^2 + \frac{(r - 1)(c - 1)}{cr} \sigma_{cr}^2.$$

We may define a statistic whose expected value involves  $\sigma_e^2$  alone by

$$V_{ncr} = M(V_{x.ns}) + M(V_{x.rsc}) - V_{x.s},$$

$$\therefore E_s(V_{ncr}) = \frac{ncr - n - cr + 1}{ncr} \sigma_e^2.$$

Our balance sheet then takes the form

Mean Sums of Squares	Divisor	Expected Value
$S_e - S$	$c - 1$	$\sigma_e^2 + n\sigma_{cr}^2 + nr\sigma_c^2$
$S_r - S$	$r - 1$	$\sigma_r^2 + n\sigma_{cr}^2 + nc\sigma_e^2$
$S_n - S$	$n - 1$	$\sigma_e^2 + cr\sigma_n^2$
$S + S_{cr} - S_c - S_r$	$(r - 1)(c - 1)$	$\sigma_e^2 + n\sigma_{cr}^2$
$S + S_q - S_n - S_{cr}$	$(cr - 1)(n - 1)$	$\sigma_e^2$

As stated, we are entitled to regard this as a balance sheet only if all 4 systematic factors are additive. We have no means of knowing this from the data of the experiment; but our concern is not to assess the contribution of a so-called factor of interaction. We wish to know whether there is a cell-effect which may or may not be indicative of interaction; and the foregoing schema shows us which statistics (*viz.* the last two) must be consistent if there is no source of variation from cell to cell within the layer other than the residual.

*Exhaustive 3-factor Analysis.* The last analysis is artificial inasmuch as we assume the knowledge that interaction between the layer-factor and the row-factor or column-factor is negligible, and its use is merely to clarify an exhaustive 3-factor analysis the aim of which is both to assess and validify the balance sheet for 3 specific criteria. Validification signifies that

we must interpret the data of the experiment with a view to demonstrating zero interaction at every level, i.e. between row and layer factors, between column and layer factors and (as above) between row and column factors. Accordingly, we postulate two hypothetical cell factors in addition to  $F_{ij.s}$ . The latter is constant within the pillar of  $n$ -cells but varies from cell to cell within the layer of  $cr$  cells and is indistinguishable from the effect of interaction between  $F_{i.cs}$  and  $F_{j.rs}$ . Similarly, we postulate  $F_{hi.s}$  as constant within the column of  $r$  cells but variable within the row-slab of  $nc$  cells; and  $F_{hj.s}$  as constant within the row of  $c$  cells but variable within the column-slab of  $nr$  cells. We denote the true variance of the three cell factors as  $\sigma_{cr}^2$  referable to  $F_{ij.s}$  as before,  $\sigma_{nc}^2$  referable to  $F_{hi.s}$  and  $\sigma_{nr}^2$  referable to  $F_{hj.s}$ . In conformity with the Model II postulates, relevant data for the construction of the balance sheet involving the 3 specific and 3 cell factors are then as in Table 5.

TABLE 5

<i>Expected Values of Components of Variance</i>							
	$e_{hij.s}$	$F_{i.cs}$	$F_{j.rs}$	$F_{h.ns}$	$F_{ij.s}$	$F_{hi.s}$	$F_{hj.s}$
$V_{x.s}$	$\frac{ncr-1}{ncr} \sigma_e^2$	$\frac{c-1}{c} \sigma_c^2$	$\frac{r-1}{r} \sigma_r^2$	$\frac{n-1}{n} \sigma_n^2$	$\frac{cr-1}{cr} \sigma_{cr}^2$	$\frac{nc-1}{nc} \sigma_{nc}^2$	$\frac{nr-1}{nr} \sigma_{nr}^2$
$V_{x.cs}$	$\frac{nr-1}{nr} \sigma_e^2$	0	$\frac{r-1}{r} \sigma_r^2$	$\frac{n-1}{n} \sigma_n^2$	$\frac{r-1}{r} \sigma_{cr}^2$	$\frac{n-1}{n} \sigma_{nc}^2$	$\frac{nr-1}{nr} \sigma_{nr}^2$
$V_{x.rs}$	$\frac{nc-1}{nc} \sigma_e^2$	$\frac{c-1}{c} \sigma_c^2$	0	$\frac{n-1}{n} \sigma_n^2$	$\frac{c-1}{c} \sigma_{cr}^2$	$\frac{nc-1}{nc} \sigma_{nc}^2$	$\frac{n-1}{n} \sigma_{nr}^2$
$V_{x.ns}$	$\frac{rc-1}{rc} \sigma_e^2$	$\frac{c-1}{c} \sigma_c^2$	$\frac{r-1}{r} \sigma_r^2$	0	$\frac{cr-1}{cr} \sigma_{cr}^2$	$\frac{c-1}{c} \sigma_{nc}^2$	$\frac{r-1}{r} \sigma_{nr}^2$
$V_{x.ncs}$	$\frac{n-1}{n} \sigma_e^2$	0	0	$\frac{n-1}{n} \sigma_n^2$	0	$\frac{n-1}{n} \sigma_{nc}^2$	$\frac{n-1}{n} \sigma_{nr}^2$
$V_{x.ncr}$	$\frac{r-1}{r} \sigma_e^2$	0	$\frac{r-1}{r} \sigma_r^2$	0	$\frac{r-1}{r} \sigma_{cr}^2$	0	$\frac{r-1}{r} \sigma_{nr}^2$
$V_{x.nrs}$	$\frac{c-1}{c} \sigma_e^2$	$\frac{c-1}{c} \sigma_c^2$	0	0	$\frac{c-1}{c} \sigma_{cr}^2$	$\frac{c-1}{c} \sigma_{nc}^2$	0

We may surmise from what has gone before that

- expected values of  $V(M_{x.cs})$ ,  $V(M_{x.rs})$  and  $V(M_{x.ns})$  will each involve the residual, one specific factor and one or more cell factors;
- expected values of  $V_{zn}$  and analogous statistics ( $V_{zc}$  and  $V_{zr}$ ) will involve only the residual and cell factors;
- a statistic whose expected value depends on the residual alone is obtainable by subtracting all the foregoing from the total sample variance ( $V_{x.s}$ )

Accordingly, we define by analogy

$$V_{zc} = M(V_{x.ns}) + M(V_{x.rs}) - M(V_{x.nrs}) - V_{x.s};$$

$$V_{zr} = M(V_{x.ns}) + M(V_{x.cs}) - M(V_{x.ncs}) - V_{x.s}.$$



## EXHAUSTIVE 3-FACTOR ANALYSIS

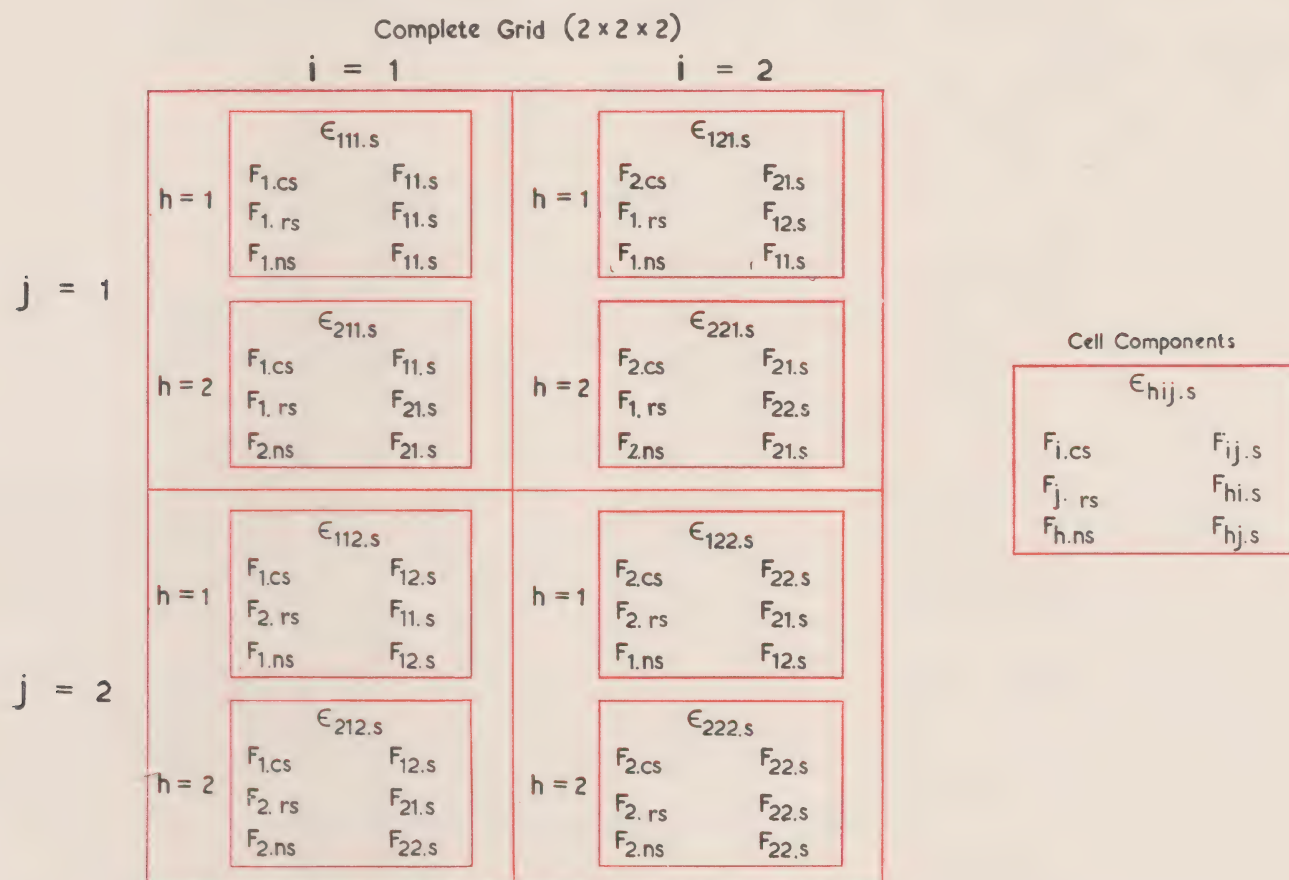


FIG. 105. The 3-factor Pattern.

Whence we derive from Table 5 the following :

$$\begin{aligned}
 E_s(V_{x.s}) &= \frac{ncr-1}{ncr} \sigma_e^2 + \frac{c-1}{c} \sigma_c^2 + \frac{r-1}{r} \sigma_r^2 + \frac{n-1}{n} \sigma_n^2 \\
 &\quad + \frac{cr-1}{cr} \sigma_{cr}^2 + \frac{nc-1}{nc} \sigma_{nc}^2 + \frac{nr-1}{nr} \sigma_{nr}^2 ; \\
 E_s \cdot V(M_{x.cs}) &= \frac{c-1}{ncr} \sigma_e^2 + \frac{c-1}{c} \sigma_c^2 + \frac{c-1}{cr} \sigma_{cr}^2 + \frac{c-1}{nc} \sigma_{nc}^2 ; \\
 E_s \cdot V(M_{x.rs}) &= \frac{r-1}{ncr} \sigma_e^2 + \frac{r-1}{r} \sigma_r^2 + \frac{r-1}{cr} \sigma_{cr}^2 + \frac{r-1}{nr} \sigma_{nr}^2 ; \\
 E_s \cdot V(M_{x.ns}) &= \frac{n-1}{ncr} \sigma_e^2 + \frac{n-1}{n} \sigma_n^2 + \frac{n-1}{nc} \sigma_{nc}^2 + \frac{n-1}{nr} \sigma_{nr}^2 ; \\
 E_s \cdot V_{zn} &= \frac{(r-1)(c-1)}{ncr} \sigma_e^2 + \frac{(r-1)(c-1)}{cr} \sigma_{cr}^2 ; \\
 E_s \cdot V_{zc} &= \frac{(r-1)(n-1)}{ncr} \sigma_e^2 + \frac{(r-1)(n-1)}{nr} \sigma_{nr}^2 ; \\
 E_s \cdot V_{zr} &= \frac{(c-1)(n-1)}{ncr} \sigma_e^2 + \frac{(c-1)(n-1)}{nc} \sigma_{nc}^2 .
 \end{aligned}$$

The expected value of the difference between the last six of the above and the first is

$$V_{ncr} = V_{w.s} - V(M_{w.cs}) - V(M_{w.rs}) - V(M_{w.ns}) - V_{zn} - V_{zc} - V_{zr}.$$

Whence we derive

$$E_s(V_{ncr}) = \frac{(n-1)(c-1)(r-1)}{ncr} \sigma_e^2.$$

#### COMPUTATION FOR EXHAUSTIVE 3-FACTOR ANALYSIS

		i = 1	i = 2	TOTAL	SUM OF SQUARES		
j = 1		$x_{11}, x_{21}$	$x_{12}, x_{22}$	$x_{11} + x_{21} + x_{12} + x_{22}$	$x_{11}^2 + x_{21}^2 + x_{12}^2 + x_{22}^2$	ncr. $V_{XCS} = S_q - S$ ncr. $V(M_{XCS}) = S_c - S$ ncr. $V(M_{XCR}) = S_r - S$ ncr. $V(M_{XCN}) = S_n - S$	ncr. $M(V_{XCS}) = S_q - S_{cr}$ ncr. $M(V_{XCR}) = S_q - S_{rc}$ ncr. $M(V_{XCN}) = S_q - S_{nr}$
j = 2		$x_{12}, x_{22}$	$x_{12}, x_{22}$	$x_{12} + x_{22} + x_{12} + x_{22}$	$x_{12}^2 + x_{22}^2 + x_{12}^2 + x_{22}^2$		
GRAND TOTAL		T.		$S_q$			
SQUARE		ncr. $S = T^2$					

		i = 1	i = 2	SQUARE TOTALS ( $T_i^2$ )		
j = 1		$x_{11}, x_{21}$	$x_{12}, x_{22}$	$(x_{11} + x_{21} + x_{12} + x_{22})^2$	GRAND TOTAL	ncr. $S_r = \sum_{i=1}^r T_i^2$
j = 2		$x_{12}, x_{22}$	$x_{12}, x_{22}$	$(x_{12} + x_{22} + x_{12} + x_{22})^2$		
GRAND TOTAL						

		h = 1	h = 2	SQUARE TOTALS ( $T_h^2$ )		
i = 1		$x_{11}, x_{12}$	$x_{21}, x_{22}$	$(x_{11} + x_{12} + x_{21} + x_{22})^2$	GRAND TOTAL	ncr. $S_c = \sum_{h=1}^c T_h^2$
i = 2		$x_{21}, x_{12}$	$x_{22}, x_{22}$	$(x_{21} + x_{12} + x_{21} + x_{22})^2$		
GRAND TOTAL						

		i = 1	i = 2	SQUARE TOTALS ( $T_i^2$ )		
h = 1		$x_{11}, x_{12}$	$x_{21}, x_{22}$	$(x_{11} + x_{12} + x_{21} + x_{22})^2$	GRAND TOTAL	ncr. $S_n = \sum_{i=1}^n T_i^2$
h = 2		$x_{21}, x_{12}$	$x_{22}, x_{22}$	$(x_{21} + x_{12} + x_{21} + x_{22})^2$		
GRAND TOTAL						

		i = 1	i = 2	TOTAL		
j = 1		$T_{11}^2 = (x_{11} + x_{21})^2$	$T_{12}^2 = (x_{12} + x_{22})^2$	$T_{11}^2 + T_{12}^2$	GRAND TOTAL	ncr. $S_{nr} = \sum_{i=1}^n \sum_{r=1}^r T_{ij}^2$
j = 2		$T_{21}^2 = (x_{21} + x_{12})^2$	$T_{22}^2 = (x_{22} + x_{22})^2$	$T_{21}^2 + T_{22}^2$		
GRAND TOTAL						

FIG. 106. Computing Schema for the 3-factor Pattern.

For computation we may reconstruct the grid of  $r$  rows,  $c$  columns and  $n$  cell entries: (a) with  $r$  rows,  $n$  columns and  $c$  cell entries with cell totals  $T_{hj}$ ; (b) with  $n$  rows  $c$  columns and  $r$  cell entries with cell totals  $T_{hi}$ . We then define by analogy with  $S_{cr}$  of (xxix) in 11.05:

$$S_{nr} = \frac{1}{c} \sum_{h=1}^n \sum_{j=1}^r T_{hj}^2; \quad S_{nc} = \frac{1}{r} \sum_{h=1}^n \sum_{i=1}^c T_{hi}^2.$$

Our complete specification of the 3-factor set-up is then:

TABLE 6

Mean Sums of Squares	Divisor	Expected Value
$S_o - S$	$c - 1$	$\sigma_o^2 + n\sigma_c^2 + n\sigma_{cr}^2 + r\sigma_{nc}^2$
$S_r - S$	$r - 1$	$\sigma_o^2 + n\sigma_r^2 + n\sigma_{cr}^2 + c\sigma_{nr}^2$
$S_n - S$	$n - 1$	$\sigma_o^2 + c\sigma_n^2 + r\sigma_{nc}^2 + c\sigma_{nr}^2$
$S + S_{cr} - S_o - S_r$	$(r-1)(c-1)$	$\sigma_o^2 + n\sigma_{cr}^2$
$S + S_{nc} - S_n - S_o$	$(n-1)(c-1)$	$\sigma_o^2 + r\sigma_{nc}^2$
$S + S_{nr} - S_n - S_r$	$(n-1)(r-1)$	$\sigma_o^2 + c\sigma_{nr}^2$
$S_o + S + S_c + S_r + S_n - S_{cr} - S_{nc} - S_{nr}$	$(n-1)(c-1)(r-1)$	$\sigma_o^2$



The last 4 entries of the foregoing table disclose which statistics must be consistent if  $\sigma_{cr}^2$ ,  $\sigma_{nc}^2$  and  $\sigma_{nr}^2$  are negligible. If they are so, we are entitled to assume that the specific components are additive. If so, and if we have also good reason to believe that the particular Model II postulates hold good, we may construct a balance sheet for the 4 remaining sources of variation in accordance with the foregoing prescription for 3-factor analysis without interaction.

### 13.07 ONE CRITERION OF CLASSIFICATION

In Chapter 7 of Vol. I we have seen that the sample mean score from a normal universe is normally distributed, as is the difference between two sample mean scores. On that basis, we can approximately assess the significance of a group mean difference. We can however formulate the null hypothesis in a different way. Given  $c$  groups of  $r$  scores, we may ask whether the assemblage of samples is homogeneous w.r.t. the column criterion of classification, i.e. that the column mean differences arise only from residual sources of variation common to all. The problem of assessing the significance of a group mean difference is the particular case, when  $c = 2$ .

If we lay out in 5 columns the heights of  $5r$  children of one sex in 5 equal age groups regardless of any peculiarities of the  $r$  score values of an age-group *inter se*, our concern is with only one criterion (*age*) of classification. On the assumption that the universe is homogeneous w.r.t. the column criterion

$$E_s(V_{x..s}) = \frac{rc - 1}{rc} \sigma^2; \quad E_s M(V_{x..cs}) = \frac{r - 1}{r} \sigma^2;$$

$$E_s V(M_{x..cs}) = E_s(V_{x..s}) - E_{s..} M(V_{x..cs}) = \frac{c - 1}{rc} \sigma^2.$$

We may define as before a statistic whose expected value is  $\sigma^2$  by the relations

$$E_s(s_c^2) = \sigma^2 \quad \text{and} \quad s_c^2 = \frac{rc}{c - 1} V(M_{x..cs}),$$

$$\therefore s_c^2 = \sum_{i=1}^{i=c} \frac{r(M_i - M)^2}{c - 1} \quad \dots \dots \dots (i)$$

Likewise we may define a second statistic whose mean value is  $\sigma^2$  by the relations

$$E_s(s_d^2) = \sigma^2 \quad \text{and} \quad s_d^2 = \frac{r}{r - 1} M(V_{x..cs}),$$

$$\therefore s_d^2 = \sum_{i=1}^{i=c} \sum_{j=1}^{j=r} \frac{(x_{ij} - M_i)^2}{c(r - 1)} \quad \dots \dots \dots (ii)$$

When our concern is with only one criterion of classification, we can develop criteria of homogeneity *without* imposing the restriction that the groups are *all of one size*. If not, we must interpret the operation  $E_r$  for extracting the mean score or mean square score derivation within the column and  $E_c$  that of extracting a mean column parameter with due regard to the weight. If there are in all  $n$  scores in the  $c$  columns, and  $r_i$  scores in the  $i$ th column, we therefore write

$$n = \sum_{i=1}^{i=c} r_i; \quad E_r(. . .) \equiv \frac{1}{r_i} \sum_{j=1}^{j=r_i} (. . .); \quad E_c(. . .) \equiv \frac{1}{n} \sum_{i=1}^{i=c} r_i(. . .).$$

As before, for the whole sample

$$E_s(V_{x..s}) = \frac{n - 1}{n} \sigma^2.$$

For the variance within the  $i$ th column, we have

$$E_s(V_{x.cs}) = \frac{r_i - 1}{r_i} \sigma^2;$$

$$E_s.M(V_{x.cs}) = E_c.E_s(V_{x.cs}) = \frac{1}{n} \sum_{i=1}^{i=c} (r_i - 1) \sigma^2 \quad . \quad . \quad . \quad (iii)$$

$$\therefore E_s.V(M_{x.cs}) = \frac{n-1}{n} \sigma^2 - \frac{1}{n} \sum_{i=1}^{i=c} (r_i - 1) \sigma^2 \quad . \quad . \quad . \quad (iv)$$

In these expressions

$$\sum_{i=1}^{i=c} (r_i - 1) = n - c,$$

$$\therefore E_s.M(V_{x.cs}) = \frac{n-c}{n} \sigma^2 \quad \text{and} \quad E_s.V(M_{x.cs}) = \frac{c-1}{n} \sigma^2 \quad . \quad . \quad . \quad (v)$$

Accordingly we define  $s_c^2$  and  $s_d^2$  by the relations

$$E_s(s_c^2) = \sigma^2 \quad \text{and} \quad s_c^2 = \frac{n}{c-1} V(M_{x.cs}) \quad . \quad . \quad . \quad (vi)$$

$$E_s(s_d^2) = \sigma^2 \quad \text{and} \quad s_d^2 = \frac{n}{n-c} M(V_{x.cs}) \quad . \quad . \quad . \quad (vii)$$

Evidently, the above are equivalent respectively to (i) and (ii) when the number of items is the same in all classes, so that  $n = rc$ . We must of course interpret  $V(M_{x.cs})$  and  $M(V_{x.cs})$  as weighted mean values, i.e.

$$V(M_{x.cs}) = \frac{1}{n} \sum_{i=1}^{i=c} r_i (M_i - M)^2 \quad \text{and} \quad M(V_{x.cs}) = \frac{1}{n} \sum_{i=1}^{i=c} \sum_{j=1}^{j=r_i} (x_{ij} - M_i)^2.$$

Thus we have

$$s_c^2 = \frac{1}{(c-1)} \sum_{i=1}^{i=c} r_i . (M_i - M)^2 \quad \text{and} \quad s_d^2 = \frac{1}{n-c} \sum_{i=1}^{i=c} \sum_{j=1}^{j=r_i} (x_{ij} - M_i)^2 \quad . \quad . \quad . \quad (viii)$$

When  $c = 2$ , we can write the column means in the form  $M_a$  and  $M_b$  referable respectively to  $r_a$  and  $r_b$  items, so that  $n = (r_a + r_b)$ . By definition therefore

$$M = \frac{r_a}{n} . M_a + \frac{r_b}{n} M_b,$$

$$\therefore (M_a - M)^2 = \frac{r_b^2}{n^2} (M_a - M_b)^2 \quad \text{and} \quad (M_b - M)^2 = \frac{r_a^2}{n^2} (M_a - M_b)^2,$$

$$\therefore s_c^2 = r_a (M_a - M)^2 + r_b (M_b - M)^2 = \frac{r_a r_b^2 + r_b r_a^2}{n^2} (M_a - M_b)^2$$

$$= \frac{r_a r_b}{r_a + r_b} (M_a - M_b)^2,$$

$$\therefore s_c^2 = \frac{(M_a - M_b)^2}{\left(\frac{1}{r_a} + \frac{1}{r_b}\right)} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (ix)$$



When  $c = 2$ , we may also write (ii) in the form

$$s_d^2 = \frac{1}{n-2} \sum_{j=1}^{j=r_a} (x_j - M_a)^2 + \frac{1}{n-2} \sum_{j=1}^{j=r_b} (x_j - M_b)^2 \quad . \quad . \quad . \quad (x)$$

We may abbreviate (x) by the substitutions

$$s_a^2 = \frac{1}{n-2} \sum_{j=1}^{j=r_a} (x_j - M_a)^2 \quad \text{and} \quad s_b^2 = \frac{1}{n-2} \sum_{j=1}^{j=r_b} (x_j - M_b)^2 \quad . \quad . \quad (xi)$$

$$\therefore s_d^2 = s_a^2 + s_b^2.$$

We shall later see (16.06) that tests of significance w.r.t. homogeneity rely on the ratios of consistent and independent estimates of the true variance ( $\sigma^2$ ) of the putative common universe. If, as we shall then see,  $s_e^2$  and  $s_d^2$  are indeed independent, a ratio appropriate to the test of significant departure from homogeneity with respect to one criterion of classification is  $s_e^2 \div s_d^2$ . When there are only two classes

$$\frac{s_e^2}{s_d^2} = \frac{(M_a - M_b)^2}{\left(\frac{1}{r_a} + \frac{1}{r_b}\right)(s_a^2 + s_b^2)} \quad . \quad . \quad . \quad . \quad (xii)$$

The relation of (xii) to the square standard score of the approximate  $c$ -test of the significance of a group mean difference (Vol. I, Chapter 7) will suggest itself at once. Our unbiased estimate of the unit sample variance based on the *pooled* data is

$$s_{ab}^2 = \frac{1}{n-1} \sum_{j=1}^{j=r} (x_j - M_{ab})^2.$$

Whence we have for the group means

$$s_1^2 = \frac{1}{r_a} s_{ab}^2 \quad \text{and} \quad s_2^2 = \frac{1}{r_b} s_{ab}^2.$$

For the variance of the mean difference we have therefore

$$\left(\frac{1}{r_a} + \frac{1}{r_b}\right) s_{ab}^2.$$

Whence the square of the appropriate  $c$ -ratio is given by

$$c^2 = \frac{(M_a - M_b)^2}{\left(\frac{1}{r_a} + \frac{1}{r_b}\right) s_{ab}^2} \quad . \quad . \quad . \quad . \quad (xiii)$$

When we have before us more than 2 groups distinguished in virtue of one criterion of classification (e.g. *breed*), our concern will commonly be to ascertain which mean values significantly differ. We must then base our estimate of the sampling error of  $M_i$  on our estimate ( $s_e^2$ ) of the residual variance  $\sigma_e^2$ . The true variance of  $M_i$  in the absence of systematic sources of variation will be ( $\sigma_e^2 \div r_i$ ) and our estimate of it will be

$$s_{m.i}^2 = \frac{s_e^2}{r_i} \quad . \quad . \quad . \quad . \quad . \quad (xiv)$$

We may regard our sample score values as divisible into two independent components, an error component and a column factor in accordance with the equation

$$x_{ij.s} = e_{ij.s} + F_{i.cs}.$$





In preceding sections of this chapter we have repeatedly met expressions exhibiting a parameter ( $p_u$ ) of a universe as the expected value of an  $n$ -fold sample parameter ( $p_s$ ) in a form which involves an integer  $f$  less than  $n$  itself, i.e.

$$E(p_s) = \frac{f}{n} \cdot p_u.$$

We then say that the sample statistic ( $p_e$ ) which is an unbiased estimate of  $p_u$  itself is

$$p_e = \frac{n}{f} \cdot p_s.$$

For the number  $f$  in such expressions, it is customary to apply the term *degrees of freedom* (d.f.). This definition does not cover all uses; and it does not obviously tie up with the literal meaning of the expression. It is therefore necessary to explain that statisticians speak of degrees of freedom to convey a seemingly quite different intention in the taxonomic domain.

The alternative usage arises from the need to distinguish between the number ( $f$ ) of score classes sufficient to define a sample and the total recorded number ( $n$ ) of such classes in contradistinction to items. In a classification involving one criterion (e.g. *suit* of cards) the rule is that  $f = (n - 1)$ . For example, it suffices to specify the number of black cards alone in a sample, if our record refers only to the binary (*black-red*) system; and we then say that the system itself has 1 d.f. If we classify our sample by suit, it is unnecessary to specify the heart score, if we have also specified the number of spades, clubs and diamonds it contains. In such a set-up  $n = 4$  and  $f = 3$ .

The student need have no misgiving if the connexion between the two uses of the expression is at this stage obscure, because each usage has a clear-cut domain. The one last mentioned may seem to be trivial at first sight. It calls for special comment only when the sample record invokes more than one criterion of classification. As a first example, we shall consider the construction of a  $2 \times 2$  table which separately assigns the numbers of red and the numbers of black cards distinguishable as *picture* and *other*. If the sample consists of  $s$  cards taken from a full pack, it suffices to know the numbers in any 3 cells of the table, since the number in the fourth cell is obtainable by subtracting the 3-cell total from  $s$ . Hence  $f = 3$ . If the sample consists of  $s_1$  cards taken from a half-pack of red only and  $s_2$  from a half-pack of black only, it suffices to know how many picture cards each contains, and  $f = 2$ .

An interesting situation arises when we classify w.r.t. one criterion of classification a sample of known size ( $s$ ) and the residual pack of  $(52 - s)$  cards, as in the schema below. If we are entitled to assume that there are 12 picture cards in the 52-card pack, our knowledge that the  $s$ -fold sample contains  $x_s$  picture cards suffices to define how many cards each cell contains and  $f = 1$ .

	Sample	Residual pack	Total
Picture	$x_s$	$x_r = (12 - x_s)$	12
Other	$s - x_s$	$40 - s + x_s$	40
Total	$s$	$52 - s$	52

On the other hand, the mere fact that the pack contains 52 cards does not necessarily mean that it is a full pack in the ordinary sense. If we cannot assume that it does contain 12 picture cards, we cannot derive the value of  $x_r$  from that of  $x_s$ , and we must assign 2 d.f. to the system.

Let us now suppose that we take two samples of known size  $s_1$  and  $s_2$  from a full pack and classify the results as in the  $3 \times 3$  table below.

	First sample	Second sample	Residue	Total known
Aces	$a_1$	$a_2$	$4 - a_1 - a_2$	4
Picture	$b_1$	$b_2$	$12 - b_1 - b_2$	12
Other	$s_1 - a_1 - b_1$	$s_2 - a_2 - b_2$	$36 - s_1 - s_2 + a_1 + a_2 + b_1 + b_2$	36
Total	$s_1$	$s_2$	$52 - s_1 - s_2$	52

In this set-up we can fill 9 cells, if we have the information to fill any 4 of them. In each of  $2 = (3 - 1)$  rows, we have to fill  $2 = (3 - 1)$  cells or  $(3 - 1)(3 - 1)$  in all. More generally, it suffices to fill  $(r - 1)(c - 1)$  cells of a grid of  $r$  rows and  $c$  columns, if we know *all* the marginal totals, but if we merely know the grand total we then need to fill  $(rc - 1)$  cells. If we do know all the marginal totals, we therefore assign  $(r - 1)(c - 1)$  d.f. to the system. If we only know the grand total  $f = (rc - 1)$ . The expression *degrees of freedom* is meaningful in this context in as far as it specifies how many cells of a grid we are *free to fill in any way consistent with the prescribed conditions without forfeiting the power to fill the remaining ones*.

We have now a clue to what lies behind the use of the term d.f. for the denominator in our unbiased sample statistics for foregoing sections of this chapter. In the two-way set-up, the total sample variance makes use of only one item of information about the  $rc$  cells in the grid, *viz.* the grand mean ( $M$ ). If this is the only *fixed* parameter of the grid, we are *free* to assign scores of any value not exceeding an aggregate of  $rcM$  to  $(rc - 1)$  cells. On the other hand, our expression for  $s_e^2$  involves also the mean value of each row and each column. In the same sense, therefore, we are free to fill only  $(c - 1)$  cells in each of  $(r - 1)$  rows. The total then is the product  $(c - 1)(r - 1)$  which replaces  $rc$  in the denominator of  $V_{e.s}$  and is what we have otherwise defined as the d.f. of the statistic  $s_e^2$ .

The foregoing remarks throw no light on the use of the term, when we later speak of a Chi-Square variate for  $f$  degrees of freedom for reasons explained in 16.04. There we shall also see why degrees of freedom are additive in the sense that the total of the divisors of a complete balance sheet is the divisor  $(rc - 1)$  or  $(ncr - 1)$ , etc. of the unbiased sample statistic of which the numerator is  $rc \cdot V_{x.s}$  or  $ncr \cdot V_{x.s}$ , etc.

#### AN ARITHMETICAL EXAMPLE

The following illustrates the procedure for estimating residual variance of data involving : (a) one criterion alone ; (b) two criteria of classification. The figures are from a paper by Rogers and Johnstone\* who used aerial slit sampling to compare the effect on the number of bacterial colonies obtained from air of a hospital ward after sweeping an oiled floor with a broom and after use of a vacuum cleaner. The same ward of a premature baby unit was swept by a broom on three successive days following the day the floor was oiled, and by a Hoover on the same three days of the next week. Counts were made on culture media exposed at intervals of one minute during the three minutes before sweeping began, at one minute intervals for four minutes while sweeping

\* Rogers and Johnstone, 1951. *J. Hyg.* **49**, 497.



went on and at one minute intervals for three minutes after it ended. Counts made on the days when the ward was swept by a broom were as follows :

<i>Successive Observations</i>	<i>Wednesday</i>	<i>Thursday</i>	<i>Friday</i>
Before sweeping 1	59	40	40
2	50	37	35
3	61	44	32
During sweeping 4	56	22	40
5	50	21	20
6	30	30	24
7	62	66	23
After sweeping 8	63	40	22
9	38	30	23
10	32	30	23

If we first assume that the only systematic source of variation is referable to the effect of sweeping, we set out our data as below. We have then 3 columns and 30 observations in all, so that  $(n - c) = (30 - 3) = 27$ .

Day and Minute	Before Sweeping ( $x_a$ )	During Sweeping ( $x_b$ )	After Sweeping ( $x_c$ )
Wednesday 1	59	56	63
2	50	50	38
3	61	30	32
4	..	62	..
Thursday 1	40	22	40
2	37	21	30
3	44	30	30
4	..	66	..
Friday 1	40	40	22
2	35	20	23
3	32	24	23
4	..	23	..
No. of observations	9	12	9

$$\begin{aligned} \sum x_a &= 398; & \sum x_b &= 444; & \sum x_c &= 301; \\ M_a &= 44.22; & M_b &= 37.0; & M_c &= 33.44; \\ \sum (x_a - M_a)^2 &= 855.56; & \sum (x_b - M_b)^2 &= 3238.0; & \sum (x_c - M_c)^2 &= 1312.22; \end{aligned}$$

$$\begin{aligned} s_e^2 &= \frac{1}{n - c} \left\{ \sum (x_a - M_a)^2 + \sum (x_b - M_b)^2 + \sum (x_c - M_c)^2 \right\} \\ &= \frac{1}{27} \{ 855.56 + 3238.0 + 1312.22 \} = 200.21. \end{aligned}$$

Thus our residual variation ( $s_e$ ) is  $\sqrt{200.21} = 14.15$  and the standard errors of the means are

$$\begin{aligned} \frac{M_a}{\sqrt{\frac{200.21}{9}}} &= 4.72 & \frac{M_b}{\sqrt{\frac{200.21}{12}}} &= 4.09 & \frac{M_c}{\sqrt{\frac{200.21}{9}}} &= 4.72. \end{aligned}$$

Thus we have for the standard errors of the differences between the means

$$(M_a - M_b) \text{ and } (M_b - M_c) \quad M_a - M_c$$

$$\sqrt{(4.72)^2 + (4.09)^2} = \pm 6.24 \quad \sqrt{2(4.72)^2} = \pm 6.67.$$

We may thus summarise the outcome of our calculations as

$$(M_a - M_b) = 7.22 \pm 6.24;$$

$$(M_a - M_c) = 10.78 \pm 6.67;$$

$$(M_b - M_c) = 3.56 \pm 6.24.$$

None of these differences is significant at the  $2\sigma$  level; but we may be misled by the circumstance that our estimate of residual variance is excessive because of failure to eliminate a second source of systematic variation, *viz.* the number of days which have elapsed since oiling the floors. Accordingly, we set out our data thus for two criteria of classification:

Minute:	Before Sweeping			During Sweeping				After Sweeping			$\sum_{i=1}^{10} x_{ir}$	$\left[ \sum_{i=1}^{10} x_{ir} \right]^2$	$M_r$
	1	2	3	4	5	6	7	8	9	10			
Wednesday	59	50	61	56	50	30	62	63	38	32	501	251001	50.1
Thursday	40	37	44	22	21	30	66	40	30	30	360	129600	36.0
Friday	40	35	32	40	20	24	23	22	23	23	282	79524	28.2
$\sum_{j=1}^3 x_{cj}$	139	122	137	118	91	84	151	125	91	85			
$\left[ \sum_{j=1}^3 x_{cj} \right]^2$	19321	14884	18769	13924	8281	7056	22801	15625	8281	7225			
$M_c$	46.33	40.67	45.67	39.33	30.33	28.00	50.33	41.67	30.33	28.33			
$\sum_{j=1}^3 x_{cj}^2$	6681	5094	6681	5220	3341	2376	8729	6053	2873	2453			

We now calculate the residual variance by the formula

$$s_e^2 = \frac{S + S_a - S_c - S_r}{(r-1)(c-1)}.$$

In this expression

$$S = \frac{1}{rc} \left[ \sum_{j=1}^r \sum_{i=1}^c x_{ij} \right]^2 = 43548.3;$$

$$S_a = \sum_{j=1}^r \sum_{i=1}^c x_{ij}^2 = 49501;$$

$$S_c = \frac{1}{r} \sum_{i=1}^c \left[ \sum_{j=1}^r x_{ij} \right]^2 = 45389.0;$$

$$S_r = \frac{1}{c} \sum_{j=1}^r \left[ \sum_{i=1}^c x_{ij} \right]^2 = 46012.5;$$

$$\therefore s_e^2 = \frac{1}{18} (1647.8) = 91.54.$$





As we shall later see, the expected values of (iii) and (iv) are respectively  $27 \div 25 = 1.08$  and  $18 \div 16 = 1.125$ . At the 5 per cent. significance level  $F$  for the divisors (so-called *degrees of freedom*) is about 4.2 and about 4.4 for 2 and 18. At the 1 per cent. level  $F$  for 2 and 18 d.f. is about 8.3. Thus there are high odds for a systematic source of variation associated with how recently the floors were oiled, and our use of the lower estimate for the residual variance is accordingly justifiable. We may accordingly recalculate the standard error of the mean for our initial lay-out of 3 columns as below :

$$\begin{array}{cc} M_a \text{ or } M_c & M_b \\ \sqrt{\frac{91.54}{9}} \simeq 3.2 & \sqrt{\frac{91.54}{12}} \simeq 2.7. \end{array}$$

Whence our corrected differences are

$$\begin{aligned} M_a - M_b &= 7.22 \pm 4.2; \\ M_a - M_c &= 10.79 \pm 4.5; \\ M_b - M_c &= 3.56 \pm 4.2. \end{aligned}$$

*Addendum.* This chapter embodies the writer's method of presenting to students of biology assumptions which underlie the *Analysis of Variance* at a time when the concept of significance had not yet become the target of a formidable body of criticism, as explained in a final chapter written after the rest of the book had gone to press. It may be too early to surmise how far the technique of estimation embodied in Churchill Eisenhart's exposition of his Model I situation will stand the test of time ; but we may be confident that the battery of test procedures based on the  $F$ -ratio, as expounded in Chapter 16 below and illustrated by the foregoing numerical example, will retain no place in a future curriculum of statistics, if the views of Wald and Neyman gain ascendancy. Meanwhile, some of the foregoing exposition may not be valueless, if it focuses attention on neglected factual assumptions implicitly invoked by those who continue to use the method.



## CHAPTER 14

# MORE ABOUT MOMENTS

### 14.01 REALITY AND RIGOUR

STATISTICS is a branch of *applied* mathematics. As such it involves issues of two kinds, correct mathematics and correct application. The theme of our last two chapters has been the latter, in so far as our concern has been to explore in what circumstances assumptions implicit in the statistical techniques known as correlation and analysis of variance are more or less relevant to the real world. Having decided, with appropriate reservations, to adopt a procedure of one or other sort, we come face to face with the problem of significance. This is partly a mathematical issue and partly a matter of common sense, i.e. awareness of reality. In Chapter 16 we shall deal with significance tests appropriate to the issues raised in Chapters 12-13. In this Chapter and the one which follows our aim will be to lay the foundations. First, we may pause with profit to get clear about what we are discussing when we invoke a significance test.

A significance test assigns the odds for or against an occurrence prescribed by a particular hypothesis. As such it tells us the ratio of all frequencies of a class of events which exclude its specification to all frequencies of a class of events which include it. From a mathematical viewpoint this is an exercise involving the summation of two sets of frequencies. From a practical viewpoint, the outcome is not unique unless we have some additional information to guide us in assessing what odds justify the rejection of the null hypothesis. Thomas Bayes pointed this out two hundred years ago, though it is still a common delusion that odds of 20 : 1 (or 370 : 1, according to taste) against the occurrence on the assumption that the null hypothesis is correct suffice to justify us in rejecting the latter.

In this context, there is no need to elaborate previous remarks (Chapter 5, Vol. I) on so prevalent a misconception. Our task is here to clarify what we do when we assess the odds ; and this involves prior knowledge of the parent universe. When we can exactly specify the universe (e.g. cubical die or card pack) from which we sample the procedure is easy to formulate and to visualise in terms of the areas of columns (Chapter 3, Vol. I) of a histogram of unit area. The nature of the sampling process and the distribution of corresponding unit scores in the parent universe supply all the information relevant to the specification of the  $r$ -fold sample by methods now familiar to the reader ; but the exact distribution of unit scores in the universe of scientific enquiry is something we rarely know. Indeed, a 2-class universe (e.g. *hearts* and *other cards*) is the only one of which it is true to say that the algebraic form, i.e.  $(q + p)^1$ , of the unit sample distribution (u.s.d.) is implicit *in the definition of it*. In the domain of representative scoring, we can so specify the algebraic form of the u.s.d. if we score by rank but then only if there are no *ties*. If so, of course, the distribution is rectangular. Otherwise, experience alone can justify whether a particular mathematical expression is or is not a reliable description of the universe of which our observations constitute a sample. In seeking such a description as a basis for a prescription of the consequences of sampling, we are not entitled to expect that we shall be able to find an exact one ; and if we have to choose between several seemingly good enough expressions, mathematical convenience will necessarily influence our choice. That is to say, we shall prefer an expression which is most amenable to the algebraic manipulations invoked by sampling theory.

One relevant consideration in this connexion is that mathematicians are much more skilful in assessing the area of segments of a continuous curve (e.g. the normal) than in summing exactly



a series of terms increasing by finite steps. We can, however, accomplish the latter approximately by visualising the contour of a histogram as the jagged outline of a continuous curve; and the outcome is often as precise as need be. Curves which are monotonic (p. 328, Vol. I) or unimodal have special descriptive advantages, because we can specify the properties of a wide range of types by the *method of moments* (Chapter 6, Vol. I). From a mathematical viewpoint, this is a great convenience; but it is important to remember that we rarely, if ever, meet a statistical universe to which we can be confident in assigning a continuous unimodal curve (e.g. the normal) as an exact description of the score distribution. It is conceivably true that such a curve would truly describe the distribution of the weights of beans in a pure line: but it is an act of faith to assert that in fact it does so. What we may know is that our observations exhibit no discontinuities if we plot them appropriately; but an inescapable limitation to any legitimate inference from this procedure is that our measuring instruments involve a scale of finite increments.

Advances in the theory of statistics leading to the distributions of Chapter 16 have gone hand in hand with greater concern for *rigour*, i.e. exacting criteria of what we can rightly infer from our *initial* assumptions; but how far this will eventually prove to be a sign of healthy growth must depend on how far we fully realise the adequacy of the initial assumptions as a description of the real world. It is especially needful to be on our guard against the plausibility of the following syllogism: (a) such and such a distribution closely describes the universe of our observations; (b) sampling from a universe with such a distribution leads to certain consequences; (c) sampling from the universe of our observations has the same consequences. This is a *non sequitur*. The first premise is purely empirical, the second purely formal and (c) is valid only if the procedure involved in (b) does not unduly magnify any of the errors implicit in the qualification *closely*.

In conformity with these considerations, we are committed to scrutinise any argument which starts with the assumption that a universe is *normal* or that of an  $r$ -fold sample therefrom is normal. We shall be in a better position to appreciate how often such an assumption is legitimate for practical purposes, if we examine the characteristics of discrete distributions with a view to formulating criteria of the adequacy of substituting a normal (or other continuous) curve as a descriptive device. In this, and in the next two chapters, we shall assume two propositions to be acceptable without elaborating earlier remarks (Chapter 6, Vol. I) on the meaning of moments as descriptive parameters of a distribution:

- (a) if all the moments of two distributions are identical, we are entitled to regard the distributions as identical;
- (b) if all the moments of one distribution ( $B$ ) lie between those of two others ( $A$  and  $C$ ) which tally sufficiently for practical purposes, we may use  $A$  or  $C$  (as most convenient) to describe  $B$ .

These assumptions sufficiently explain the need for a closer study of procedures for evaluating moments (14.02 and 14.03) and for examining (as in other sections of this chapter) the circumstances in which the moments of a prescribed sampling distribution approach those of the normal or other continuous function which is easy to tabulate for reference. Not all the material in what follows is essential to an understanding of the rest of the book; and *the reader may well prefer to scan this chapter quickly, returning to it at a later stage if necessary*.

When we say that a continuous variate ( $y$ ) is satisfactory as a function descriptive of a discrete sampling distribution, we may adopt either or both of two criteria, having in mind the relation of the *approximate* fitting curve  $y = f(x)$  to the contour of the *exact* histogram. We



construct the latter \* on the assumption that the area of each column of height  $y_x$  on a base extending from  $(x - \frac{1}{2}\Delta x)$  to  $(x + \frac{1}{2}\Delta x)$  is numerically equivalent to the frequency ( $f_x$ ) of the score  $x$ . Since the sum of all frequencies is *unity* by definition, for a score with positive values only

$$\sum_{x=0}^{x=\infty} f_x = 1 = \sum_{x=0}^{x=\infty} y_x \cdot \Delta x \quad . \quad . \quad . \quad . \quad (i)$$

The reader will note that the exact limits are irrelevant, since  $f_x = 0$  if  $x$  lies outside the range of permissible score values. If  $x$  may be negative or positive, we must write the above as

$$\sum_{x=-\infty}^{x=\infty} f_x = 1 = \sum_{-1}^{\infty} y_x \cdot \Delta x \quad . \quad . \quad . \quad . \quad (\text{ii})$$

The sum of the frequencies in the range  $x = a$  to  $x = b$  is

$$\sum_{x=a}^{x=b} f_x = S_{ab} = \sum_{x=a}^{x=b} y_x \cdot \Delta x \quad . \quad . \quad . \quad . \quad (\text{iii})$$

The actual boundaries of the corresponding columns of the histogram of unit area are not the mid-points  $a$  and  $b$  but  $(a - \frac{1}{2}\Delta x)$  and  $(b + \frac{1}{2}\Delta x)$ .

With due regard to these considerations, we may ask for a fitting curve such that

- (a) if the ordinate  $y$  at  $x$  goes nearly through the mid-point of the upper extremity of the corresponding column of the histogram, we can approximately specify the frequency of the score  $x$  if we know the scale  $\Delta x$ , i.e.

$$y \cong y_x \quad \text{and} \quad y . \Delta x \cong f_x \qquad \cdot \qquad \cdot \qquad \cdot \qquad \cdot \qquad \cdot \qquad (\text{iv})$$

- (b) if the area bounded by the ordinates of the curve at  $x = a$  and  $x = b$  corresponds closely with that of the segment of the histogram with the corresponding boundaries at  $(a - \frac{1}{2}\Delta x)$  and  $(b + \frac{1}{2}\Delta x)$ , we can approximately specify the net frequencies of score values within the range, i.e.

$$\sum_{x=a}^{x=b} y_x \cdot \Delta x = \int_{a-\frac{1}{2}\Delta x}^{b+\frac{1}{2}\Delta x} y \cdot dx = I_{ab} . \quad . \quad . \quad . \quad . \quad (\text{v})$$

In accordance with our definition of frequency, any such fitting curve must also fulfil the criterion of unit area over the whole range of permissible values of  $x$  itself : (i) positive only, (ii) negative and positive :

$$(i) \int_0^{\infty} y \cdot dx = 1 \quad (ii) \int_{-\infty}^{\infty} y \cdot dx = 1 \quad . \quad . \quad . \quad (vi)$$

To make the foregoing criteria explicit we have to define a criterion of error, as follows :

- (a) if  $e_c$  is the error consistent with a satisfactory *ordinate* fit in a specified range, we may express (iv) more explicitly as

[illegible]

- (b) if  $e_{gh}$  is the error consistent with a satisfactory *expectation* fit

$$S_{ab} - I_{ab} < |e_{ab}| \quad . \quad . \quad . \quad . \quad . \quad . \quad (\text{viii})$$

\* The reader may find it useful in this context to read more detailed remarks on the build-up of the frequency histogram in 15.01 below.

In discussions on curve fitting, there is not always an explicit recognition that the two criteria defined by (vii) and (viii) have different domains of relevance. If we are seeking a criterion of goodness of fit for an *empirical* distribution (vii) may have more to commend it, inasmuch as our main object in so *graduating* our data may be to assign a numerical value to the frequency of the occurrence of a particular score. This is not the end in view when we perform a significance test. Our concern is then to sum frequencies of all score values which do not exceed a specified limit or limits; and (viii) is the appropriate criterion. The distinction is important in so far as no unimodal descriptive function could be satisfactory in the sense defined by (vii), if the exact distribution has gaps such as we have seen (Chapter 4, Vol. I) to be characteristic of the distribution of the proportionate score difference distribution of co-prime samples from an infinite 2-class universe. On the other hand, the fact that we can eliminate such gaps by grouping disposes of any objection against seeking a satisfactory fit in accordance with (viii).

#### 14.02 MOMENT GENERATING FUNCTION

The type of generating function dealt with in 11.08 is one of many ways of summarising the operations of the independence grid with a view to specification of a sum or difference distribution. The essential desideratum of such a g.f. is that the dummy factor which identifies the score value associated with a particular frequency as a co-factor obeys the product rule for indices, *viz.*  $t^a \cdot t^b = t^{a+b}$  and  $t^a \cdot t^{-b} = t^{a-b}$ . This, of course, is equally true if we express  $t$  in exponential form as  $t = e^h$ , so that  $e^{ah} \cdot e^{bh} = e^{(a+b)h}$ . If  $u_x$  is the frequency of the score  $x$  we may therefore label our dummy co-factor as  $e^{xh}$  defining a new class of generating functions, exactly as before with the substitution  $t = e^h$ , so that

$$G_u = u_0 + u_1 e^h + u_2 e^{2h} + u_3 e^{3h} \dots u_n e^{nh} \quad . \quad . \quad . \quad (i)$$

$$G_{-u} = u_0 + u_1 e^{-h} + u_2 e^{-2h} + u_3 e^{-3h} \dots u_n e^{-nh} \quad . \quad . \quad . \quad (ii)$$

If we label the cell frequencies of the grid for 2 independent variates  $u_x$  and  $v_x$  as  $y_{ij} = u_i v_j$  in accordance with the schemata exhibited in 11.08, the rules for the summation of frequencies of the sum ( $s$ ) and difference ( $d$ ) are as there given, *viz.* :

$$Y_s = \sum_{x=0}^{x=s} y_{(s-x)x} \quad \text{and} \quad Y_d = \sum_{x=0}^{x=d} y_{(d+x)x}.$$

If  $G_u$  and  $G_{-u}$  are each referable to score frequencies of the *unit sample distribution*,  $x_a$  and  $x_b$  are respectively score-sums of independent  $a$ -fold and  $b$ -fold samples from the same universe,

$$G(x_a + x_b) = G_u^{a+b} \quad \text{and} \quad G(x_a - x_b) = G_u^a \cdot G_{-u}^b \quad . \quad . \quad . \quad (iii)$$

For the proportionate (or mean) score difference distribution :

$$G\left(\frac{x_a}{a} - \frac{x_b}{b}\right) = (u_0 + u_1 e^{\frac{h}{a}} + u_2 e^{\frac{2h}{a}} + u_3 e^{\frac{3h}{a}} \dots)^a (u_0 + u_1 e^{-\frac{h}{b}} + u_2 e^{-\frac{2h}{b}} + u_3 e^{-\frac{3h}{b}} \dots)^b.$$



TABLE 1

THE M.G.F. AS A GRID OPERATION

Second Zero Moment of Heart-Score sums w.r.t. 3-fold and 2-fold samples taken with replacement from two full packs

3-fold Sample

Raw-Score	0	1	2	3
Exponential Score	$e^0$	$e^h$	$e^{2h}$	$e^{3h}$
Frequency	$\frac{27}{64}$	$\frac{27}{64}$	$\frac{9}{64}$	$\frac{1}{64}$

2-fold Sample

0 $e^0$ $\frac{9}{16}$	$s = 0; s^2 = 0$ $e^{sh} = e^0$ $y_{00} = \frac{243}{1024}$	$s = 1; s^2 = 1$ $e^{sh} = e^h$ $y_{10} = \frac{243}{1024}$	$s = 2; s^2 = 4$ $e^{sh} = e^{2h}$ $y_{20} = \frac{81}{1024}$	$s = 3; s^2 = 9$ $e^{sh} = e^{3h}$ $y_{30} = \frac{9}{1024}$
1 $e^h$ $\frac{6}{16}$	$s = 1; s^2 = 1$ $e^{sh} = e^h$ $y_{01} = \frac{182}{1024}$	$s = 2; s^2 = 4$ $e^{sh} = e^{2h}$ $y_{11} = \frac{182}{1024}$	$s = 3; s^2 = 9$ $e^{sh} = e^{3h}$ $y_{21} = \frac{54}{1024}$	$s = 4; s^2 = 16$ $e^{sh} = e^{4h}$ $y_{31} = \frac{6}{1024}$
2 $e^{2h}$ $\frac{1}{16}$	$s = 2; s^2 = 4$ $e^{sh} = e^{2h}$ $y_{02} = \frac{27}{1024}$	$s = 3; s^2 = 9$ $e^{sh} = e^{3h}$ $y_{12} = \frac{27}{1024}$	$s = 4; s^2 = 16$ $e^{sh} = e^{4h}$ $y_{22} = \frac{9}{1024}$	$s = 5; s^2 = 25$ $e^{sh} = e^{5h}$ $y_{32} = \frac{1}{1024}$

Diagonal Frequencies ( $\times 1024$ )

$Y_0$	$Y_1$	$Y_2$	$Y_3$	$Y_4$	$Y_5$
243	405	270	90	15	1

$$E(e^{sh}) = 2^{-10}(243 + 405e^h + 270e^{2h} + 90e^{3h} + 15e^{4h} + e^{5h});$$

$$D_h^2 E(e^{sh}) = 2^{-10}(405e^h + 1080e^{2h} + 810e^{3h} + 240e^{4h} + 25e^{5h});$$

$$D_{h=0}^2 E(e^{sh}) = 2^{-10}(405 + 1080 + 810 + 240 + 25) = \frac{5}{2};$$

$$E(s_2) = \mu_2(s) = \frac{5}{2} = 5pq + 5^2p^2.$$

In doing this we have done nothing new, since  $t = e^h$  is merely a dummy for purposes of identification; but the substitution has the merit that the generating function so expressed has a dual purpose. We can use it as before to write down the terms of a sampling distribution; but we can also use it to evaluate its moments in a new way by taking advantage of the differential property of  $e^x$ , viz.:

$$e^{xh} = 1 + xh + \frac{x^2h^2}{2!} + \frac{x^3h^3}{3!} + \frac{x^4h^4}{4!} + \frac{x^5h^5}{5!} + \frac{x^6h^6}{6!} \text{ etc.};$$

$$D_h(e^{xh}) = 0 + x + x^2h + \frac{x^3h^2}{2!} + \frac{x^4h^3}{3!} + \frac{x^5h^4}{4!} + \frac{x^6h^5}{5!} \text{ etc.};$$

$$D_h^2(e^{xh}) = 0 + 0 + x^2 + x^3h + \frac{x^4h^2}{2!} + \frac{x^5h^3}{3!} + \frac{x^6h^4}{4!} \text{ etc.};$$

$$D_h^3(e^{xh}) = 0 + 0 + 0 + x^3 + x^4h + \frac{x^5h^2}{2!} + \frac{x^6h^3}{3!} \text{ etc.}$$

TABLE 2

THE M.G.F. AS A GRID OPERATION

Second Zero Moment of Heart-Score difference ( $d$ ) w.r.t. 3-fold and 2-fold samples taken with replacement from two full packs

3-fold Sample				
Raw-Score	0	1	2	3
Exponential Score	$e^0$	$e^h$	$e^{2h}$	$e^{3h}$
Frequency	$\frac{27}{64}$	$\frac{27}{64}$	$\frac{9}{64}$	$\frac{1}{64}$

2-fold Sample				
0	$e^0$	$\frac{9}{16}$	$d = 0; d^2 = 0$ $e^{dh} = e^0$	$d = 1; d^2 = 1$ $e^{dh} = e^h$
			$y_{00} = \frac{243}{1024}$	$y_{10} = \frac{243}{1024}$
1	$e^{-h}$	$\frac{6}{16}$	$d = -1; d^2 = 1$ $e^{dh} = e^{-h}$	$d = 0; d^2 = 0$ $e^{dh} = e^0$
			$y_{01} = \frac{162}{1024}$	$y_{11} = \frac{162}{1024}$
2	$e^{-2h}$	$\frac{1}{16}$	$d = -2; d^2 = 4$ $e^{dh} = e^{-2h}$	$d = -1; d^2 = 1$ $e^{dh} = e^{-h}$
			$y_{02} = \frac{27}{1024}$	$y_{12} = \frac{27}{1024}$
			$d = 0; d^2 = 0$ $e^{dh} = e^0$	$d = 1; d^2 = 1$ $e^{dh} = e^h$
			$y_{20} = \frac{81}{1024}$	$y_{30} = \frac{9}{1024}$
			$d = 1; d^2 = 1$ $e^{dh} = e^h$	$d = 2; d^2 = 4$ $e^{dh} = e^{2h}$
			$y_{21} = \frac{54}{1024}$	$y_{31} = \frac{6}{1024}$
			$d = 2; d^2 = 4$ $e^{dh} = e^{2h}$	$d = 3; d^2 = 9$ $e^{dh} = e^{3h}$
			$y_{22} = \frac{9}{1024}$	$y_{32} = \frac{1}{1024}$

Diagonal Frequencies ( $\times 1024$ )

$Y_{-2}$	$Y_{-1}$	$Y_0$	$Y_1$	$Y_2$	$Y_3$
27	189	414	298	87	9

$$E(e^{dh}) = 2^{-10}(27e^{-2h} + 189e^{-h} + 414 + 298e^h + 87e^{2h} + 9e^{3h});$$

$$D_h^2 \cdot E(e^{dh}) = 2^{-10}(108e^{-2h} + 189e^{-h} + 298e^h + 348e^{2h} + 81e^{3h});$$

$$D_{h=0}^2 E(e^{dh}) = 2^{-10}(108 + 189 + 298 + 348 + 81) = 1 = \mu_2(d);$$

$$E(d^2) = 2^{-10}(0 \cdot 414 + 1 \cdot 189 + 298 + 4 \cdot 27 + 87 + 9 \cdot 9) = 1 = \mu_2(d).$$



$$D_h^k(e^{xh}) = x^k + x^{k+1}h + x^{k+2}\frac{h^2}{2!} + x^{k+3}\frac{h^3}{3!} \text{ etc.}$$
$$D_h^k(u_x \cdot e^{xh})_{h=0} = u_x \cdot x^k;$$

$$\therefore D_h^k(G_u)_{h=0} = \sum_0^\infty u_x \cdot x^k.$$

$$\mu_k = \sum_{x=0}^{x=\infty} u_x \cdot x^k;$$

$$\therefore \mu_k = D_h^k(G_u)_{h=0} \quad . \quad . \quad . \quad . \quad . \quad (\text{iv})$$

*Example 1.*—Find the moments of the distribution of the mean score for the 3-fold toss of a tetrahedral die with face scores 1, 3, 5, 7. The g.f. of the u.s.d. is

$$\frac{1}{4}(e^{\hbar} + e^{3\hbar} + e^{5\hbar} + e^{7\hbar}) = \frac{e^{\hbar}}{4}(1 + e^{2\hbar} + e^{4\hbar} + e^{6\hbar}).$$

$$\frac{e^{3\hbar}}{64}(1 + e^{2\hbar} + e^{4\hbar} + e^{6\hbar})^3 = \frac{e^{3\hbar}}{64}(1 + 3e^{2\hbar} + 6e^{4\hbar} + 10e^{6\hbar} + 12e^{8\hbar} + 12e^{10\hbar} + 10e^{12\hbar} + 6e^{14\hbar} + 3e^{16\hbar} + e^{18\hbar}).$$
$$\frac{1}{64}(e^h + 3e^{\frac{5h}{3}} + 6e^{\frac{7h}{3}} + 10e^{\frac{9h}{3}} + 12e^{\frac{11h}{3}} \dots e^{\frac{21h}{3}}) = G.$$
$$(D_h^k \cdot e^{\alpha h})_{h=0} = x^k,$$

$$\therefore \mu_k = \frac{1}{64} [1 + 3\left(\frac{5}{3}\right)^k + 6\left(\frac{7}{3}\right)^k + 10\left(\frac{9}{3}\right)^k \dots]$$

\* \* \* \* \*

\* We here assume that the range of scores is positive only. More generally,  $\mu_k = E(x^k)$  for a score  $x$  with weighted summation over the whole range from  $-\infty$  to  $+\infty$  as for the *urette universe* of 14.05 and exercise below. If  $M_x = 0$  we then have  $\mu_k = m_k$ .

What is more important is that we can use it to derive results of general interest. In this connexion we may note that the operation of determining the  $k$ th mean moment ( $m_k$ ) is precisely analogous to the foregoing procedure. If  $y_X$  is the frequency of the score deviation  $X$  our m.g.f. is

$$G_u(X) = \sum_{-\infty}^{\infty} y_X \cdot e^{Xh}.$$

The  $k$ th differential w.r.t.  $h$  is

$$\sum_{-\infty}^{\infty} y_X (X^k + h \cdot X^{k+1} + \frac{h^2}{2!} X^{k+2} + \frac{h^3}{3!} X^{k+3} \dots \text{etc.}).$$

When  $h = 0$  this reduces to

$$\sum_{-\infty}^{\infty} y_X X^k = E(X^k) = m_k.$$

Expressions of this type simplify in virtue of the identity  $m_1 = 0 = E(X)$ . *En passant* we may also note that  $\mu_0 = E(x^0) = 1 = E(X^0) = m_0$ , i.e.  $G_u(x) = 1 = G_u(X)$  when  $h = 0$ . An alternative way of labelling our frequencies leads to a familiar series of formulae. If we write  $\mu_1 = M$ , so that  $X = x - M$ :

$$\begin{aligned} \sum_{-\infty}^{\infty} y_X \cdot e^{Xh} &= \sum_0^{\infty} y_x \cdot e^{(x-M)h} = e^{-Mh} \sum_0^{\infty} y_x \cdot e^{xh}; \\ \therefore G_u(X) &= e^{-Mh} G_u(x). \end{aligned}$$

If we write for brevity  $w = G_u(X)$  and  $z = G_u(x)$ ,

$$\begin{aligned} D_h w &= e^{-Mh} \cdot D_h z - M e^{-Mh} z; \\ D_h^2 w &= e^{-Mh} \cdot D_h^2 z - M e^{-Mh} \cdot D_h z + M^2 e^{-Mh} z - M e^{-Mh} \cdot D_h z; \\ D_h^3 w &= e^{-Mh} \cdot D_h^3 z - 3M e^{-Mh} \cdot D_h^2 z + 3M^2 e^{-Mh} \cdot D_h z - M^3 e^{-Mh} z. \end{aligned}$$

When  $h = 0$  so that  $z = 1 = e^{-Mh}$  and  $D_h^k z = \mu_k$ , we have

$$m_1 = 0; m_2 = \mu_2 - M^2; m_3 = \mu_3 - 3M\mu_2 + 2M^3.$$

Similarly, we may obtain higher mean moments in terms of zero moments. Likewise, we obtain the familiar expressions for zero moments in terms of mean moments, if we write

$$G_u(x) = e^{Mh} G_u(X).$$

In what follows we assume that our scores increase by unit steps ( $\Delta x = 1$ ), since the appropriate scalar transformation is straightforward if this is not so. If  $z = ax$ , so that  $\Delta z = a$  we may write

$$\begin{aligned} E(z^k) &= E(x \Delta z)^k; \\ \therefore \mu_k(z) &= (\Delta z)^k \cdot \mu_k(x). \end{aligned}$$

In particular, if  $x_s$  is the score-sum of the  $p$ -fold sample, that of the mean score is  $(x_s \div p) = x_m$  and  $x_s = p \cdot x_m$ , whence  $\mu_k(x_s) = p^k \cdot \mu_k(x_m)$  and

$$\mu_k(x_m) = \frac{\mu_k(x_s)}{p^k}.$$



$$z^r = (u_0 + u_1 e^h + u_2 e^{2h} + u_3 e^{3h} \dots)^r.$$
$$D_h(D_h z)^2 = 2D_h z \cdot D_h^2 z \quad \text{and} \quad D_h(D_h z)^3 = 3(D_h z)^2 \cdot D_h^2 z.$$
$$D_h z^r = r z^{r-1} D_h z;$$

$$D_h^2 z^r = r^{(2)} z^{r-2} (D_h z)^2 + r z^{r-2} D_h^2 z;$$

$$D_h^3 z^r = r^{(3)} z^{r-3} (D_h z)^3 + 3r^{(2)} z^{r-2} D_h z \cdot D_h^2 z + r z^{r-1} D_h^3 z;$$

$$D_h^4 z^r = r^{(4)} z^{r-4} (D_h z)^4 + 6r^{(3)} z^{r-3} (D_h z)^2 D_h^2 z + 3r^{(2)} z^{r-2} (D_h^2 z)^2 + 4r^{(2)} z^{r-2} D_h z \cdot D_h^3 z + r z^{r-1} D_h^4 z.$$

$$\mu_1(r) = r\mu_1 \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (v)$$

[illegible]

$$\mu_3(r) = r^{(3)}\mu_1^3 + 3r^{(2)}\mu_1 \cdot \mu_2 + r\mu_3 \quad . \quad . \quad . \quad . \quad . \quad . \quad (\text{vii})$$

$$\mu_4(r) = r^{(4)}\mu_1^4 + 6r^{(3)}\mu_1^2 \cdot \mu_2 + 3r^{(2)}\mu_2^2 + 4r^{(2)}\mu_1 \cdot \mu_3 + r\mu_4 \quad . \quad (\text{viii})$$

$$m_1(r) = 0 = m_1 \quad . \quad . \quad . \quad . \quad . \quad . \quad (\text{ix})$$

$$m_9(r) = rm_2 \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (X)$$

$$m_3(r) = rm_3 \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (\text{xi})$$

$$m_4(r) = 3r^{(2)}m_2^2 + rm_4 \quad . \quad . \quad . \quad . \quad . \quad (xii)$$

$$\mu_k = p; m_k = q(-p)^k + pq^k.$$
$$\mu_1(r) = rp;$$

$$\mu_2(r) = r(r-1)p^2 + rp = r^2p^2 + rp(1-p);$$

$$\mu_3(r) = r^{(3)}p^3 + 3r^{(2)}p^2 + rp;$$

$$\mu_4(r) = r^{(4)}p^4 + 6r^{(3)}p^3 + 7r^{(2)}p^2 + rp.$$

Similarly, we derive

$$\begin{aligned} m_2(r) &= rpq; \\ m_3(r) &= rpq(q^2 - p^2) = rpq(q - p); \\ m_4(r) &= 3r(r - 1)p^2q^2 + rpq(1 - 3pq) \\ &= 3r^2p^2q^2 + rpq(1 - 6pq). \end{aligned}$$

Another result we may usefully invoke at a later stage is the distribution of the difference between independent unit samples from the same universe. We may write the g.f. of the zero moment of the u.s.d. as

$$u = \sum_{x=0}^{x=\infty} u_x \cdot e^{xh}; \quad (D_h^k u)_{h=0} = \mu_k.$$

Since we may regard the paired score difference as the sum of positive and negative scores distributed in the same way, we may write for the corresponding g.f. of the negative score distribution

$$v = \sum_{x=0}^{x=\infty} u_x \cdot e^{-xh}; \quad (D_h^k v)_{h=0} = (-1)^k \mu_k.$$

The m.g.f. of the difference distribution in accordance with the product rule is then  $uv$ , and

$$\mu_k(d) = (D_h^k \cdot uv)_{h=0}.$$

By successive differentiation we have

$$\begin{aligned} D_h uv &= u D_h v + v D_h u; \\ D_h^2 uv &= u D_h^2 v + 2 D_h u \cdot D_h v + v D_h^2 u; \\ D_h^3 uv &= u D_h^3 v + 3 D_h u \cdot D_h^2 v + 3 D_h^2 u \cdot D_h v + v D_h^3 u; \\ D_h^4 uv &= u D_h^4 v + 4 D_h u \cdot D_h^3 v + 6 D_h^2 u \cdot D_h^2 v + 4 D_h^3 u \cdot D_h v + v D_h^4 u. \end{aligned}$$

In these expressions when  $h = 0$

$$\begin{aligned} u = 1 = v; \quad D_h u = \mu_1 = -D_h v; \quad D_h^2 u = \mu_2 = D_h^2 v; \\ D_h^3 u = \mu_3 = -D_h^3 v; \quad D_h^4 u = \mu_4 = D_h^4 v. \end{aligned}$$

We therefore obtain

$$\mu_1(d) = 0 = \mu_3(d); \quad \mu_2(d) = 2\mu_1; \quad \mu_4(d) = 2\mu_4 + 6\mu_2^2.$$

Since the mean of the distribution is zero, the mean moments and zero moments are identical. So we have

$$m_2(d) = 2m_2; \quad m_3(d) = 0; \quad m_4(d) = 2m_4 + 6m_2^2.$$

More generally, we see by the same procedure that the odd moments of the paired difference distribution from the same universe are all zero and the even moments are the same as those of the 2-fold sample. If the u.s.d. is symmetrical, the odd moments of the 2-fold sample score-sum distribution are also zero, as is true of a universe if successive terms of the expansion  $(\frac{1}{2} + \frac{1}{2})^a$  define the score frequencies. As we have then seen, the frequencies of the 2-fold sample score-sum in the range  $0, 1, 2 \dots 2a$  then tally with successive terms of  $(\frac{1}{2} + \frac{1}{2})^{2a}$ . We thus arrive at the following conclusion: if the binomial  $(\frac{1}{2} + \frac{1}{2})^a$  defines the u.s.d. in the range 0 to  $a$ , the



binomial  $(\frac{1}{2} + \frac{1}{2})^{2a}$  defines the distribution of both the pair-score difference with range  $-a$  to  $+a$  and of the pair-score sum with range 0 to  $2a$ . If the scores of the u.s.d. increase by  $\Delta x$  from  $m$  to  $m + a\Delta x$ , those of the pair-difference increase by  $\Delta x$  from  $-a\Delta x$  to  $+a\Delta x$  and those of the pair-score sum from  $2m$  to  $2m + 2a\Delta x$  with the same increment.

Throughout this section our concern has been with discrete distributions. We may extend our scope, if we recall remarks at the end of 14.01. The frequency ( $f_x$ ) of a score  $x$  is numerically equivalent to the corresponding area of the histogram of unit area, being  $y_x \Delta x$  if  $y_x$  is the height and  $\Delta x$  the base whose midpoint is  $x$ . We may thus speak of  $f_x = y_x \cdot \Delta x$  as the frequency of a score in the range  $(x \pm \frac{1}{2}\Delta x)$ . When  $\Delta x$  becomes indefinitely small, we may write  $y = y_x$  as the ordinate at  $x$  and  $y \cdot dx$  as the frequency of the score  $x$  in the range  $x \pm \frac{1}{2}dx$ . We then have

$$\begin{aligned}\sum_0^\infty f_x &= \int_0^\infty y \cdot dx; \\ \sum_0^\infty f_x \cdot x^k &= \int_0^\infty y \cdot x^k \cdot dx = \mu_k; \\ \sum_0^\infty f_x \cdot e^{xh} &= \int_0^\infty y \cdot e^{xh} \cdot dx.\end{aligned}$$

The last expression is the generating function of the zero moments, when all score values are positive. If  $Y \cdot dX$  is the frequency of the score deviation  $X$  in the interval  $X \pm \frac{1}{2}dX$ , the g.f. of the mean moments is

$$\int_{-\infty}^{\infty} Y \cdot e^{xh} \cdot dX.$$

When the function  $Y$  is symmetrical we may write this as

$$2 \int_0^\infty Y \cdot e^{xh} \cdot dX.$$

Such expressions are of no use for determining moments unless they are integrable, as is true when the distribution is normal (*vide* 14.04 below). A class of continuous distributions of special importance is that of Gamma variates specified as such in accordance with the definition of the Gamma function, *viz.* :

$$\int_0^\infty e^{-kx} \cdot x^{n-1} \cdot dx = \frac{\Gamma(n)}{k^n} \text{ so that } \frac{k^n}{\Gamma(n)} \int_0^\infty e^{-kx} \cdot x^{n-1} \cdot dx = 1.$$

Thus the ordinate equation of a Gamma variate is

$$y = \frac{k^n \cdot e^{-kx} \cdot x^{n-1}}{\Gamma(n)}.$$

The g.f. of the zero moments is therefore

$$\begin{aligned}\frac{k^n}{\Gamma(n)} \int_0^\infty e^{-kx} \cdot x^{n-1} \cdot e^{xh} \cdot dx &= \frac{k^n}{\Gamma(n)} \int_0^\infty e^{-(k-h)x} \cdot x^{n-1} \cdot dx; \\ \therefore G_u &= \frac{k^n}{\Gamma(n)} \cdot \frac{\Gamma(n)}{(k-h)^n} = k^n (k-h)^{-n}.\end{aligned}$$

In this expression we may expand  $(k - h)^{-n}$  by the binomial theorem, *viz.* :

$$(k - h)^{-n} = k^{-n} + nk^{-n-1}h + \frac{n(n+1)}{2!}k^{-n-2}h^2 + \frac{n(n+1)(n+2)}{3!}k^{-n-3}h^3 \dots \text{etc.},$$

$$\therefore G_u = \sum_{r=0}^{\infty} \frac{(n+r-1)^{(r)}}{r!} k^{-r} h^r;$$

$$D_h G_u = nk^{-1} + (n+1)^{(2)}k^{-2}h + \frac{(n+2)^{(3)}k^{-3}h^2}{2!} + \frac{(n+3)^{(4)}k^{-4}h^3}{3!} \text{ etc.};$$

$$D_h^2 G_u = (n+1)^{(2)}k^{-2} + (n+2)^{(3)}k^{-3}h + \frac{(n+3)^{(4)}k^{-4}h^2}{2!} \dots \text{etc.};$$

$$D_h^3 G_u = (n+2)^{(3)}k^{-3} + (n+3)^{(4)}k^{-4}h + \frac{(n+4)^{(5)}k^{-5}h^2}{2!} \dots \text{etc.}$$

When  $h = 0$  we thus get

$$\mu_1 = nk^{-1}; \mu_2 = (n+1)^{(2)}k^{-2}; \mu_3 = (n+2)^{(3)}k^{-3} \text{ etc.}$$

We shall obtain this result by another method in Chapter 15.

#### EXERCISE 14.02

1. Find the first 4 zero and mean moments of the mean score distribution of the 3-fold toss of a tetrahedral die with face scores

(i) 1, 2, 3, 4.

(ii) 3, 6, 9, 12.

(iii) 1, 2, 2, 3.

(iv) 2, 7, 7, 12.

2. Find the same moments for the distribution of both the raw-score and proportionate-score difference w.r.t. 3-fold and 2-fold tosses of each of the above.

3. Investigate the first 4 mean moments of the distribution of the mean score of samples from a 3-class universe with the following u.s.d. scores:  $-1, 0, +1$ , frequencies  $p_a, p_b, p_c$ .

4. Find general expressions for the first six Pearson coefficients ( $\beta_1$  to  $\beta_6$ ) of the u.s.d. of the above for the symmetrical case when  $p_a = p_c$  (for definition of  $\beta_3$  to  $\beta_6$ , see 14.04 below).

5. Find expressions for the 6th and 8th mean moments of the 6-fold sample from the symmetrical 3-class universe and hence for the corresponding values of  $\beta_4$  and  $\beta_6$  (see 14.04).

6. Derive expressions for the first 8 moments of the score difference distribution of 6-fold samples from the symmetrical 3-class universe and show that they are equal to the moments of the distribution of the sum of the difference between 6 pairs of unit samples.

7. Tabulate results of the above for  $p_a = p_c = \frac{2}{5}, \frac{1}{3}, \frac{1}{4}, \frac{1}{6}, \frac{1}{10}$  and check all results by using the probability generating functions of the distribution (see Ex. 6, 11.09).

8. Show that the m.g.f. of a Poisson system (*vide* Chapter 10 of Vol. I) for  $r$  unit samples from  $r$  2-class universes is

$$G(\mu) = \prod_{i=1}^{i=r} (q_i + p_i e^{\mu_i}).$$

Check this result numerically for samples of 3 by putting  $p_1 = \frac{1}{2}, p_2 = \frac{2}{3}, p_3 = \frac{3}{4}$ .



9. If we have two series of urns with parameters  $p_{b.i}$  and  $p_{a.i}$  definitive of the u.s.d., show that the m.g.f. of the mean of  $r$  paired differences in sampling with replacement is

$$G(\mu) = \prod_{i=1}^{i=r} (A_i e^{-t} + B_i + C_i e^t)$$

in which expression

$$A_i = p_{a.i} \cdot q_{b.i}; \quad B_i = p_{a.i} \cdot p_{b.i} + q_{a.i} \cdot q_{b.i}; \quad C_i = p_{b.i} \cdot q_{a.i}.$$

10. For the last set-up cite the value of  $\beta_2$  for the 3-fold sample of paired differences when  $p_{b.i} = k \cdot p_{a.i}$ .

### 14.03 FACTORIAL MOMENTS

We can sometimes sidestep difficulties w.r.t. derivation or computation of moments, more especially moments of a discrete distribution, by recourse to analogous indices in which factorials take the place of ordinary powers of the score  $x$  or of its deviation  $X = (x - M)$  from the mean. Thus we define zero and mean *factorial moments* as follows :

$$\mu_{(k)} = \sum y_x \cdot x^{(k)} \quad \text{and} \quad m_{(k)} = \sum Y_X \cdot X^{(k)} \quad . \quad . \quad . \quad (i)$$

Here we concern ourselves solely with the zero factorial moments, recalling that

$$x^{(k)} = x(x-1)(x-2) \dots (x-k+1) = \frac{x!}{(x-k)!} \quad . \quad . \quad . \quad (ii)$$

For the first 8 powers we have

$$\begin{aligned} x^{(1)} &= x; \\ x^{(2)} &= x^2 - x; \\ x^{(3)} &= x^3 - 3x^2 + 2x; \\ x^{(4)} &= x^4 - 6x^3 + 11x^2 - 6x; \\ x^{(5)} &= x^5 - 10x^4 + 35x^3 - 50x^2 + 24x; \\ x^{(6)} &= x^6 - 15x^5 + 85x^4 - 225x^3 + 274x^2 - 120x; \\ x^{(7)} &= x^7 - 21x^6 + 175x^5 - 735x^4 + 1624x^3 - 1764x^2 + 720x; \\ x^{(8)} &= x^8 - 28x^7 + 322x^6 - 1960x^5 + 6769x^4 - 13132x^3 + 13068x^2 - 5040x. \end{aligned}$$

Since we can always express  $x^{(k)}$  in the form

$$\begin{aligned} x^{(k)} &= {}^kK_0 x^k + {}^kK_1 x^{k-1} + {}^kK_2 x^{k-2} \dots {}^kK_{k-1} x \\ &= \sum_{r=0}^{r=k-1} {}^kK_r \cdot x^{k-r}; \\ \therefore \mu_{(k)} &= E[x^{(k)}] = \sum_{r=0}^{r=k-1} {}^kK_r E(x^{k-r}) \\ &= \sum_{r=0}^{r=k-1} {}^kK_r \mu_{k-r}. \end{aligned}$$

From the relations cited above

$$\begin{aligned} \mu_{(1)} &= E(x) = \mu_1; \\ \mu_{(2)} &= E(x^2) - E(x) = \mu_2 - \mu_1; \\ \mu_{(3)} &= E(x^3) - 3E(x^2) + 2E(x) = \mu_3 - 3\mu_2 + 2\mu_1 \text{ etc.} \end{aligned}$$

We can thus derive any ordinary moment as a series of zero factorial moments, the first eight being

$$\begin{aligned}\mu_1 &= \mu_{(1)}; \\ \mu_2 &= \mu_{(2)} + \mu_{(1)}; \\ \mu_3 &= \mu_{(3)} + 3\mu_{(2)} + \mu_{(1)}; \\ \mu_4 &= \mu_{(4)} + 6\mu_{(3)} + 7\mu_{(2)} + \mu_{(1)}; \\ \mu_5 &= \mu_{(5)} + 10\mu_{(4)} + 25\mu_{(3)} + 15\mu_{(2)} + \mu_{(1)}; \\ \mu_6 &= \mu_{(6)} + 15\mu_{(5)} + 65\mu_{(4)} + 90\mu_{(3)} + 31\mu_{(2)} + \mu_{(1)}; \\ \mu_7 &= \mu_{(7)} + 21\mu_{(6)} + 140\mu_{(5)} + 350\mu_{(4)} + 301\mu_{(3)} + 63\mu_{(2)} + \mu_{(1)}; \\ \mu_8 &= \mu_{(8)} + 28\mu_{(7)} + 266\mu_{(6)} + 1050\mu_{(5)} + 1701\mu_{(4)} + 966\mu_{(3)} + 127\mu_{(2)} + \mu_{(1)}.\end{aligned}$$

The advantage of this arises from the fact that simple general expressions for factorial moments may be obtainable, when it is not possible to derive simple expressions for moments of the more familiar sort. The following examples will show that this is so:

(a) *The Poisson Distribution*

For this distribution with mean  $M$  and range  $x = 0$  to  $\infty$

$$y_x = \frac{e^{-M} M^x}{x!},$$

$$\therefore \mu_{(k)} = e^{-M} \sum_0^{\infty} \frac{M^x x^{(k)}}{x!} \quad \dots \dots \dots \quad \text{(iii)}$$

In virtue of (ii) we may write

$$\frac{M^x x^{(k)}}{x!} = \frac{M^x}{(x-k)!} = \frac{M^k \cdot M^{x-k}}{(x-k)!},$$

$$\therefore \mu_{(k)} = e^{-M} \cdot M^k \sum_0^{\infty} \frac{M^{x-k}}{(x-k)!} \quad \dots \dots \dots \quad \text{(iv)}$$

All terms in the summation vanish if  $x < k$ , so that we are concerned only with  $x = k, k+1, k+2 \dots \infty$ , i.e.

$$\begin{aligned}\sum_0^{\infty} \frac{M^{x-k}}{(x-k)!} &= \frac{M^0}{0!} + \frac{M^1}{1!} + \frac{M^2}{2!} + \frac{M^3}{3!} \dots \\ &= 1 + M + \frac{M^2}{2!} + \frac{M^3}{3!} \dots \\ &= e^M.\end{aligned}$$

Whence by substitution in (iv)

$$\mu_{(k)} = M^k \quad \dots \dots \dots \quad \text{(v)}$$

Hence we derive

$$\begin{aligned}\mu_1 &= \mu_{(1)} = M; \\ \mu_2 &= \mu_{(2)} + \mu_{(1)} = M^2 + M; \\ \mu_3 &= \mu_{(3)} + 3\mu_{(2)} + \mu_{(1)} = M^3 + 3M^2 + M; \\ \mu_4 &= \mu_{(4)} + 6\mu_{(3)} + 7\mu_{(2)} + \mu_{(1)} = M^4 + 6M^3 + 7M^2 + M.\end{aligned}$$



The mean moments are obtainable from the zero moments by recourse to the now familiar formulae, so that

$$m_2 = M = m_3; \quad m_4 = M(1 + 3M);$$

$$\therefore \beta_1 = \frac{1}{M} \quad \text{and} \quad \beta_2 = 3 + \frac{1}{M}. \quad \text{. . . . . (vi)}$$

For the Poisson distribution we thus have the relation

$$\beta_2 = 3 + \beta_1 \quad \text{. . . . . (vii)}$$

The reader will recall (6.08 in Vol. I) that the Type III distribution is more leptokurtic for the same measure of skewness, since

$$\beta_2 = 3 + \frac{3}{2}\beta_1 \quad \text{. . . . . (viii)}$$

By proceeding in the same way we derive, for the higher Pearson coefficients (p. 599),

$$\beta_3 = \frac{10}{M} + \frac{1}{M^2}; \quad \beta_4 = 15 + \frac{25}{M} + \frac{1}{M^2} \quad \text{. . . . . (ix)}$$

$$\beta_5 = \frac{105}{M} + \frac{56}{M^2} + \frac{1}{M^3}; \quad \beta_6 = 105 + \frac{490}{M} + \frac{119}{M^2} + \frac{1}{M^3}.$$

In Chapter 5 of Vol. I, we have seen that the Poisson approaches the normal distribution when  $M > 9$ . It is instructive to compare the Pearson coefficients of the normal with those of the Poisson distribution when  $M = 1, 10$  and  $100$ .

		$\beta_1$	$\beta_2$	$\beta_3$	$\beta_4$	$\beta_5$	$\beta_6$
Poisson	$M = 1$	1.0	4.0	11.0	41.0	162.0	715.0
„	$M = 10$	0.1	3.1	1.01	17.51	11.06	155.191
„	$M = 100$	0.01	3.01	0.1001	15.2501	1.0556	109.91
Normal		0.0	3.0	0.0	15.0	0.0	105.0

### (b) The Rectangular Distribution

If the universe is  $n$ -fold with scores  $1, 2, 3 \dots n$ ,

$$\mu_{(k)} = \sum_{x=1}^{x=n} \frac{x^{(k)}}{n} \quad \text{. . . . . (x)}$$

By (v) in 11.07

$$\sum_{x=1}^{x=n} x^{(k)} = \frac{(n+1)^{(k+1)}}{(k+1)} = \frac{(n+1)n(n-1)^{(k-1)}}{k+1};$$

$$\therefore \mu_{(k)} = \frac{(n+1)(n-1)^{(k-1)}}{(k+1)} \quad \text{. . . . . (xi)}$$

Evidently,

$$\mu_{(1)} = \frac{n+1}{2}; \quad \mu_{(2)} = \frac{n^2-1}{3};$$

$$\mu_{(3)} = \frac{(n^2-1)(n-2)}{4}; \quad \mu_{(4)} = \frac{(n^2-1)(n^2-5n+6)}{5}, \text{ etc.}$$

From these, and by means of the next four of the series, we derive the following values for the ordinary zero and mean moments :

$$\begin{aligned}
 \mu_2 &= \frac{(n+1)(2n+1)}{6} & m_2 &= \frac{n^2-1}{12} \\
 \mu_3 &= \frac{n(n+1)^2}{4} & m_3 &= 0 \\
 \mu_4 &= \frac{(n+1)(2n+1)(3n^2+3n-1)}{30} & m_4 &= \frac{(n^2-1)(3n^2-7)}{240} \\
 \mu_5 &= \frac{n(n+1)^2(2n^2+2n-1)}{12} & m_5 &= 0 \\
 \mu_6 &= \frac{(n+1)(2n+1)(3n^4+6n^3-3n+1)}{42} & m_6 &= \frac{(n^2-1)(3n^4-18n^2+31)}{1344} \\
 \mu_7 &= \frac{n(n+1)^2(3n^4+6n^3-n^2-4n+2)}{24} & m_7 &= 0 \\
 \mu_8 &= \frac{(n+1)(2n+1)(5n^6+15n^5+5n^4-15n^3-n^2+9n-3)}{90} & m_8 &= \frac{(n^2-1)(5n^6-55n^4+239n^2-381)}{11520}
 \end{aligned}$$

Hence we derive the following exact expressions for Pearson coefficients of *even* order, all odd ones being of zero value :

$$\beta_2 = \frac{9}{5} - \frac{12}{5(n^2-1)}; \quad \beta_4 = \frac{27}{7} - \frac{108}{7(n^2-1)} + \frac{144}{7(n^2-1)^2} \quad \text{. . . . . (xii)}$$

$$\beta_6 = 9 - \frac{72}{n^2-1} + \frac{1296}{5(n^2-1)^2} - \frac{1728}{5(n^2-1)^3} \quad \text{. . . . . (xiii)}$$

### (c) *A Binomial Distribution*

If the definitive binomial of the raw score ( $x$ ) distribution is  $(p+q)^r$

$$\mu_{(k)} = \sum_{x=0}^{x=r} x^{(k)} r_{(x)} p^x q^{r-x} \quad \text{. . . . . (xiv)}$$

The derivation of an appropriate expression is easy by successive partial differentiation of the binomial series. Thus

$$\begin{aligned}
 \frac{\partial^2}{\partial p^2} \sum_0^r r_{(x)} p^x q^{r-x} &= \frac{\partial^2}{\partial p^2} (p+q)^r; \\
 \therefore \sum_0^r x(x-1) r_{(x)} p^{x-2} q^{r-x} &= r(r-1)(p+q)^{r-2}.
 \end{aligned}$$

More generally we thus derive

$$\begin{aligned}
 \frac{\partial^k}{\partial p^k} \sum_0^r r_{(x)} p^x q^{r-x} &= \frac{\partial^k}{\partial p^k} (p+q)^r; \\
 \therefore \sum_0^r x^{(k)} r_{(x)} p^{x-k} q^{r-x} &= r^{(k)} (p+q)^{r-k} = r^{(k)}; \\
 \therefore \sum_0^r x^{(k)} r_{(x)} p^x q^{r-x} &= r^{(k)} p^k.
 \end{aligned}$$



Alternatively, we may proceed as follows :

$$\begin{aligned} r_{(x)} p^x q^{r-x} x^{(k)} &= \frac{r^{(k)}(r-k)!}{x!(r-x)!} p^k p^{x-k} q^{r-x} \frac{x!}{(x-k)!} \\ &= r^{(k)} p^k \frac{(r-k)!}{(r-x)!(x-k)!} p^{x-k} q^{r-x}. \end{aligned}$$

If we put  $u = (x - k)$  and  $v = (r - k)$  in the above so that  $(r - x) = (v - u)$

$$\sum_{x=0}^{x=r} r_{(x)} p^x q^{r-x} x^{(k)} = r^{(k)} p^k \sum_{u=-k}^{u=v} \frac{v!}{u!(v-u)!} p^u q^{v-u}.$$

Since  $u!$  is infinite when  $u$  is negative, all terms in the range  $u = -1$  to  $u = -k$  vanish, and

$$\begin{aligned} \sum_{x=0}^{x=r} r_{(x)} p^x q^{r-x} x^{(k)} &= r^{(k)} p^k \sum_{u=0}^{u=v} \frac{v!}{u!(v-u)!} p^u q^{v-u} = r^{(k)} p^k (p+q)^v, \\ \therefore \sum_{x=0}^{x=r} r_{(x)} p^x q^{r-x} x^{(k)} &= r^{(k)} p^k, \\ \therefore \mu_{(k)} &= r^{(k)} p^k. \end{aligned} \quad \text{(xv)}$$

Hence we derive the following expressions for the zero moments :

$$\begin{aligned} \mu_2 &= r^2 p^2 + r p q; \\ \mu_3 &= r^3 p^3 + 3 r^2 p^2 q + r p q (q - p); \\ \mu_4 &= r^4 p^4 + 6 r^3 p^3 q + r^2 p^2 q (7 - 11 p) + r p q (1 - 6 p q). \end{aligned}$$

The corresponding mean moments are

$$\begin{aligned} m_2 &= r p q; \\ m_3 &= r p q (q - p); \\ m_4 &= 3 r^2 p^2 q^2 + r p q (1 - 6 p q). \end{aligned} \quad \text{(xvi)}$$

We derive in the same way the higher even moments, e.g.

$$m_6 = 15 r^3 p^3 q^3 + 5 r^2 p^2 q^2 (5 - 26 p q) + r p q (1 - 30 p q + 120 p^2 q^2) \quad \text{(xvii)}$$

#### EXERCISE 14.03

1. Determine the first 4 factorial moments of the unit, 2-fold, 3-fold and 4-fold sample distributions of an infinite universe of score values 1, 2, 3 (a) in the ratio 1 : 2 : 1 ; (b) in the ratio 1 : 4 : 1.

2. From the foregoing results obtain the first four ordinary moments about the mean, and  $\beta_1$  and  $\beta_2$  for each distribution.

3. Repeat the foregoing exercises for score values of  $-1, 0$  and  $+1$  in the same ratios. Compare the results.

#### 14.04 THE NORMAL DISTRIBUTION

The only tabulated sample distribution we have used in Vol. I as a sufficiently satisfactory descriptive function in the sense defined in 14.01 is the *normal*. When we seek to establish in what sense the normal integral provides a satisfactory fit to an exactly definable distribution which involves recourse to laborious computation, it is important to be clear about the significance

$$c = \frac{X}{\sqrt{V}} = \frac{X}{\sigma} \quad . \quad . \quad . \quad . \quad . \quad . \quad (i)$$
$$V_c = E\left(\frac{X}{\sigma}\right)^2 = \frac{1}{\sigma^2}E(X^2);$$

$$c = \frac{(x - rp)}{\sqrt{rpq}}.$$

$p$	$r$	$c$
$\frac{1}{2}$	16	$\frac{(x-8)}{2}$
$\frac{1}{4}$	36	$\frac{2(x-9)}{3\sqrt{3}}$
$\frac{1}{5}$	25	$\frac{(x-5)}{2}$
$\frac{1}{10}$	100	$\frac{(x-10)}{3}$

If our only object is to compare frequencies, no other precaution calls for comment. However, our main concern in statistics is with summation of frequencies ; and we use the table of the normal integral defining the area under the curve from  $-\infty$  (or zero) up to the assigned significance level. We then have to remember to make the half interval correction of (v) in 14.01, as explained more fully in Chapter 3 of Vol. I (p. 112). If the score increases by unit steps ( $\Delta x = 1$ ), the boundaries of the base of the histogram enclosing an area which corresponds to the sum of frequencies of score deviations from  $X = a$  up to  $X = b$  are respectively at  $(a - \frac{1}{2})$  and  $(b + \frac{1}{2})$ . Thus the appropriate  $X$  co-ordinates of the fitting curve which approximately goes through the midpoint of the uppermost margin of the histogram columns are for  $X$  and  $-X$  respectively  $(X + \frac{1}{2})$  and  $-(X + \frac{1}{2})$ . More generally, if the score increases by  $\Delta x$  the boundaries of the base of the histogram enclosing an area which corresponds to the sum of



score deviations from  $X = a$  up to  $X = b$  are at  $(a - \frac{1}{2}\Delta x)$  and  $(b + \frac{1}{2}\Delta x)$ . The appropriate  $X$  co-ordinates of the fitting curve for  $X$  and  $-X$  respectively are then  $(X + \frac{1}{2}\Delta x)$  and  $-(X + \frac{1}{2}\Delta x)$ .

To express our score in standard form referable to a mean  $M$  and variance  $\sigma^2$  with due regard to this correction, we must distinguish three cases :

(i) to sum frequencies from  $-X$  to  $\infty$  or from  $-\infty$  to  $-X$ , we write

$$c = \frac{-(x - M + \frac{1}{2}\Delta x)}{\sigma};$$

(ii) to sum frequencies from  $-\infty$  to  $X$  or from  $X$  to  $\infty$

$$c = \frac{(x - M + \frac{1}{2}\Delta x)}{\sigma};$$

(iii) to sum frequencies in the range  $\pm X$ , we write

$$c = \frac{\pm (x - M + \frac{1}{2}\Delta x)}{\sigma}.$$

*Score-sum Distribution of Unit Variance.* In Chapter 16 we shall have to make extensive use of a property of score distributions so defined that the variance is unity ; and we may appropriately dispose of it in this context. Suppose a player records as his score  $w$  the sum of the individual results  $x_a, x_b, \dots, x_n$  of single trials from  $n$  identical dice weighted by a particular constant ( $a, b, \dots, n$ ) which need not be positive, i.e.

$$w = a \cdot x_a + b \cdot x_b + c \cdot x_c \dots n \cdot x_n.$$

This is, of course, equivalent to renumbering the score faces by an appropriate scalar change  $z_a = a \cdot x_a, z_b = b \cdot x_b$ , so that

$$w = z_a + z_b + z_c \dots z_n.$$

The effect of this scalar change is, of course,

$$V(z_a) = a^2 V(x_a); V(z_b) = b^2 V(x_b), \text{ etc.}$$

In virtue of independence, we may write the variance of the score-sum distribution as

$$\begin{aligned} V_w &= V(z_a) + V(z_b) + V(z_c) \dots + V(z_n) \\ &= a^2 V(x_a) + b^2 V(x_b) + c^2 V(x_c) \dots n^2 V(x_n). \end{aligned}$$

If the dice are identical  $V(x_a) = V(x_b) = V(x_c)$ , etc.  $= V_x$ , and

$$V_w = (a^2 + b^2 + c^2 \dots n^2) V_x \quad \dots \dots \dots \quad (\text{iii})$$

In this expression  $V_x$  is the variance of the unit sample (single toss) distribution of the die, and we make the variance of the player's score distribution identical with it if the rule of the game prescribes that

$$a^2 + b^2 + c^2 \dots n^2 = 1 \quad \dots \dots \dots \quad (\text{iv})$$

If we vary the rule of the game by prescribing that the player records the weighted average of the *standard scores* of the  $n$  trials

$$w = a \cdot c_a + b \cdot c_b \dots n \cdot c_n.$$

It follows from (iii) that

$$V_w = a^2 + b^2 + \dots n^2 \quad \dots \dots \dots \quad (\text{v})$$





More generally, we define higher Pearson moments of *even* order as even moments of the standard score distribution, e.g.

$$\beta_4 = \frac{m_6}{m_2^3} = m_6(c); \quad \beta_6 = \frac{m_8}{m_2^4} = m_8(c) \quad . \quad . \quad . \quad . \quad (viii)$$

Pearson's first coefficient of *odd* order is the square of the third moment of the standard score distribution, being

$$\beta_1 = \frac{m_3^2}{m_2^3} = m_3^2(c).$$

Squaring makes it irrelevant whether the mode lies right or left of the mean. The definition of the next two coefficients of odd order is

$$\beta_3 = \frac{m_3 m_5}{m_2^4} = m_3(c) \cdot m_5(c); \quad \beta_5 = \frac{m_3 m_7}{m_2^5} = m_3(c) \cdot m_7(c) \quad . \quad . \quad . \quad . \quad (ix)$$

In the light of previous remarks, we thus see that Pearson's coefficients embody a method of comparison of distributions eliminating distortion arising from difference of scale. The symbolism he introduced is confusing for more than one reason. It would be preferable to label coefficients of even order so that the subscript would be that of the corresponding moment of the standard score distribution; and the definition of coefficients of odd order is doubly exceptionable. By defining  $\beta_1$  as the square of the third moment of the standard score distribution we eliminate the sign of the skewness, so that one distribution otherwise identical with another may be its mirror image; and the definition of higher coefficients of odd order has a disadvantage aside from the fact that it is inconsistent with the pattern of the coefficients of even order. Though they must all vanish if the distribution is symmetrical, the converse is not true. They must vanish if  $m_3 = 0$ , and this is consistent with the possibility that the distribution is *not* symmetrical, since higher moments of odd order need not then vanish.

For the normal distribution it suffices to define the values of Pearson's coefficients of even order, since all moments of odd order must vanish in virtue of symmetry. We shall therefore write

$$m_{2k}(c) = E(c^{2k}) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}c^2} \cdot c^{2k} \cdot dc.$$

Since the  $c$ -distribution is symmetrical

$$m_{2k}(c) = \frac{2}{\sqrt{2\pi}} \int_0^{\infty} e^{-\frac{1}{2}c^2} \cdot c^{2k} \cdot dc.$$

To evaluate this integral, we now substitute  $c^2 = C$  so that

$$dc = \frac{dc}{dC} \cdot dC = \frac{1}{2} C^{-\frac{1}{2}} \cdot dC,$$

$$\begin{aligned} \therefore m_{2k}(c) &= \frac{1}{\sqrt{2\pi}} \int_0^{\infty} e^{-\frac{1}{2}C} \cdot C^{k-\frac{1}{2}} \cdot dC \\ &= \frac{(\frac{1}{2})^{\frac{1}{2}}}{\Gamma(\frac{1}{2})} \int_0^{\infty} e^{-\frac{1}{2}C} \cdot C^{(k+\frac{1}{2})-1} \cdot dC \quad . \quad . \quad . \quad . \quad (x) \end{aligned}$$

The integral in the last expression is a Gamma function, whose value is by (xi) in 6.05 (Vol. I)

$$\frac{\Gamma(k + \frac{1}{2})}{(\frac{1}{2})^{k+\frac{1}{2}}};$$

$$\therefore m_{2k}(c) = \frac{2^k \Gamma(k + \frac{1}{2})}{\Gamma(\frac{1}{2})} \quad \text{. . . . . (xi)}$$

We have already seen (p. 249, Vol. I) that :

$$\Gamma(1\frac{1}{2}) = \frac{1}{2} \Gamma(\frac{1}{2}); \quad \Gamma(2\frac{1}{2}) = \frac{3}{2} \cdot \frac{1}{2} \Gamma(\frac{1}{2});$$

$$\Gamma(3\frac{1}{2}) = \frac{5}{2} \cdot \frac{3}{2} \cdot \frac{1}{2} \Gamma(\frac{1}{2}); \quad \Gamma(4\frac{1}{2}) = \frac{7}{2} \cdot \frac{5}{2} \cdot \frac{3}{2} \cdot \frac{1}{2} \Gamma(\frac{1}{2}).$$

In general therefore

$$\Gamma(k + \frac{1}{2}) = \frac{1 \cdot 3 \cdot 5 \cdot \dots (2k - 1)}{2^k} \Gamma(\frac{1}{2}).$$

By substitution in (xi) we thus derive

$$m_{2k}(c) = 1 \cdot 3 \cdot 5 \cdot \dots (2k - 1).$$

By setting  $k = 1, 2$ , etc. in the above we therefore derive

$$m_2(c) = 1; \quad m_4(c) = 3; \quad m_6(c) = 15; \quad m_8(c) = 105.$$

Hence from (vii) and (viii)

$$\beta_2 = 3; \quad \beta_4 = 15; \quad \beta_6 = 105 \quad \text{. . . . . (xii)}$$

*The Square Normal Standard Score.* We may here pause to notice that the foregoing examination of the moments of the distribution of the normal score of unit variance leads at once to the derivation of the zero moments of the distribution of its square ( $C = c^2$ ). By definition the  $k$ th zero moment of the  $C$ -distribution is

$$\mu_k(c^2) = E(C^k) = E(c^{2k}),$$

$$\therefore \mu_k(c^2) = m_{2k}(c) \quad \text{. . . . . (xiii)}$$

$$\therefore \mu_1(C) = 1; \quad \mu_2(C) = 3; \quad \mu_3(C) = 15; \quad \mu_4(C) = 105, \text{ etc.}$$

We shall have occasion to examine the meaning of the results last stated more fully in 15.02. Meanwhile, the reader may with profit recall the Type III variate (p. 257, Vol. I) whose p.d. equation is

$$f(C) = \frac{(\frac{1}{2})^{\frac{1}{2}}}{\Gamma(\frac{1}{2})} e^{-\frac{1}{2}C} C^{\frac{1}{2}-1} \quad \text{. . . . . (xiv)}$$

The  $k$ th zero moment of the above distribution is

$$\mu_k(C) = \frac{(\frac{1}{2})^{\frac{1}{2}}}{\Gamma(\frac{1}{2})} \int_0^\infty e^{-\frac{1}{2}C} C^{-\frac{1}{2}} \cdot C^k dC,$$

$$\therefore \mu_k(C) = \frac{(\frac{1}{2})^{\frac{1}{2}}}{\Gamma(\frac{1}{2})} \int_0^\infty e^{-\frac{1}{2}C} C^{(k+\frac{1}{2})-1} \cdot dC.$$

The above is identical with (x), whence

$$\mu_k(C) = m_{2k}(c).$$

We may therefore write

$$\mu_k(C) = \mu_k(c^2) \quad \text{. . . . . (xv)}$$

Thus (xiv) defines the distribution whose moments of any order are identical with those of the distribution of a square normal standard score, as we shall establish by a different procedure in 15.02 below.



*M.G.F. of the Normal Distribution.* From (vii)-(viii) above, we easily obtain the moments of the normal distribution when the variance is not unity. We may express (vii)-(viii) and (xii) in the form

$$\beta_{2k} = \frac{m_{2k+2}}{V^{k+1}} \quad \text{and} \quad \beta_{2k} = (2k+1)(2k-1)(2k-3) \dots 5.3.1.$$

Whence by putting  $k = 0, 1, 2$ , etc. we have the following results :

$$m_2 = V; \quad m_4 = 3V^2; \quad m_6 = 15V^3; \quad m_8 = 108V^4 \quad \dots \quad \dots \quad \dots \quad (xvi)$$

We shall later make use of the moment generating function to derive certain useful properties of normal distributions. Before we derive its form, we may show that the following function does in fact generate the moments defined by (xvi) :

$$G_u = e^{\frac{1}{2}Vh^2} \quad \dots \quad \dots \quad \dots \quad \dots \quad \dots \quad (xvii)$$

By expansion in the usual way

$$G_u = 1 + \frac{Vh^2}{2} + \frac{V^2h^4}{4 \cdot 2!} + \frac{V^3h^6}{8 \cdot 3!} + \frac{V^4h^8}{16 \cdot 4!} + \frac{V^5h^{10}}{32 \cdot 5!} \dots \dots \text{etc.}$$

Whence we obtain

$$D_h G_u = Vh + \frac{V^2h^3}{2} + \frac{V^3h^5}{8} + \frac{V^4h^7}{48} + \frac{V^5h^9}{384} \dots \dots \text{etc.};$$

$$D_h^2 G_u = V + \frac{3V^2h^2}{2} + \frac{5V^3h^4}{8} + \frac{7V^4h^6}{48} \dots \dots \text{etc.};$$

$$D_h^3 G_u = 3V^2h + \frac{5V^3h^3}{2} + \frac{7V^4h^5}{8} \dots \dots \text{etc.};$$

$$D_h^4 G_u = 3V^2 + \frac{15V^3h^2}{2} + \frac{35V^4h^4}{8} \dots \dots \text{etc.};$$

$$D_h^5 G_u = 15V^3h + \frac{35V^4h^3}{2} \dots \dots \text{etc.};$$

$$D_h^6 G_u = 15V^3 + \frac{105V^4h^2}{2} \dots \dots \text{etc.}$$

Whence we have

$$\begin{aligned} (D_h \cdot G_u)_{h=0} &= 0; & (D_h^2 \cdot G_u)_{h=0} &= V; & (D_h^3 \cdot G_u)_{h=0} &= 0; \\ (D_h^4 \cdot G_u)_{h=0} &= 3V^2; & (D_h^5 \cdot G_u)_{h=0} &= 0; & (D_h^6 \cdot G_u)_{h=0} &= 15V^3; \\ (D_h^7 \cdot G_u)_{h=0} &= 0; & (D_h^8 \cdot G_u)_{h=0} &= 105V^4 \text{ etc.} \end{aligned}$$

The derivation of (xvii) is as follows. By definition,

$$\begin{aligned} G_u &= \frac{1}{\sqrt{2\pi V}} \int_{-\infty}^{\infty} e^{-\frac{X^2}{2V}} \cdot e^{Xh} \cdot dX \\ &= \frac{1}{\sqrt{2\pi V}} \int_{-\infty}^{\infty} \exp - \frac{(X^2 - 2VhX + V^2h^2)}{2V} e^{\frac{1}{2}Vh^2} \cdot dX \\ &= \frac{e^{\frac{1}{2}Vh^2}}{\sqrt{2\pi V}} \int_{-\infty}^{\infty} \exp - \frac{(X - Vh)^2}{2V} dX. \end{aligned}$$

If we put  $(X - Vh) = z$ , so that  $dz = dX$

$$G_u = \frac{e^{\frac{1}{2}Vh^2}}{\sqrt{2\pi V}} \int_{-\infty}^{\infty} e^{-\frac{z^2}{2V}} dz.$$

In this expression

$$\frac{1}{\sqrt{2\pi V}} \int_{-\infty}^{\infty} e^{-\frac{z^2}{2V}} dz = 1,$$

$$\therefore G_u = e^{\frac{1}{2}Vh^2}.$$

Thus the m.g.f. of the score-sum (and hence of the mean score with appropriate change of scale) of  $r$  unit samples taken from the same normal universe is

$$G_u^r = e^{\frac{1}{2}rVh^2}.$$

This is the m.g.f. of a normal distribution of variance  $rV$ . That of the mean score is the m.g.f. of a normal distribution of variance  $rV \div r^2 = V \div r$ . The m.g.f. of the score-sum of two samples from normal universes with u.s.d. variance  $V_a$  and  $V_b$  is

$$e^{\frac{1}{2}V_a h^2} \cdot e^{\frac{1}{2}V_b h^2} = e^{\frac{1}{2}(V_a + V_b)h^2}.$$

This is the m.g.f. of a normal distribution of variance  $(V_a + V_b)$ . In virtue of symmetry (11.08, p. 473) it is also that of the distribution of the score difference. We may sum up these results as follows :

- (i) If the u.s.d. of a universe is normal with variance  $V$ , the distribution of the score-sum of the  $a$ -fold sample is normal with variance  $aV$  and that of the mean score is normal with variance  $V \div a$  ;
- (ii) the distribution of the sum and difference of  $a$ -fold and  $b$ -fold samples from the same normal universe is normal with variance  $(a + b)V = V_a + V_b$ .

#### 14.05 MOMENTS OF THE DISTRIBUTION OF THE MEAN

In Chapter 7 of Vol. I we have seen that it is always possible to express in the form  $V_t = t \cdot V_u$  the variance ( $V_t$ ) of the random  $t$ -fold sample distribution of the score-sum in terms of the variance  $V_u$  of the unit sample distribution. We shall now employ results obtained in 11.02–11.03 to determine the value of higher mean moments of the distribution of the score-sum (and hence of the mean score) of  $t$ -fold samples in terms of the mean moments of the unit sampling distribution of the parent universe ; and hence to establish an important conclusion with reference to the distribution of sample means, already foreshadowed in 14.02. Our assumptions are : (i) that the samples are random, i.e. that choice of one item is *independent* of that of another ; (ii) that none of the moments concerned is infinite. The last condition is true of any distribution of *discrete* scores. In virtue of (i), the relevant formulae are those derived by application of the product rule. So we can derive the ensuing results by recourse to generating functions, as indicated in 14.02. Here we shall do so by a more elementary procedure.

It will be convenient to denote the  $k$ th mean moment of the unit sampling distribution by  $m_k$  and that of the  $t$ -fold score-sum by  $m_k(t)$ . We now recall definitions given in 11.03, and write the 3rd moment of the distribution of the score-sum of the  $(a + b)$ -fold sample in terms of those of the  $a$ -fold and  $b$ -fold samples as

$$\begin{aligned} m_3(a + b) &= E(\overline{a + b - M_a + M_b})^3 = E(\overline{a - M_a} + \overline{b - M_b})^3 \\ &= E(a - M_a)^3 + 3E(a - M_a)^2(b - M_b) + 3E(a - M_a)(b - M_b)^2 + E(b - M_b)^3 \\ &= m_3(a) + 3m_2(a) \cdot m_1(b) + 3m_1(a) \cdot m_2(b) + m_3(b). \end{aligned}$$



$$m_3(a + b) = m_3(a) + m_3(b),$$

$$\therefore m_3(t + 1) = m_3(t) + m_3.$$
$$m_3(t) = t \cdot m_3.$$
$$\begin{aligned} m_4(t+1) &= m_4(t) + 4m_3(t) \cdot m_1 + 6m_2(t) \cdot m_2 + 4m_1(t) \cdot m_3 + m_4 \\ &= m_4(t) + 6m_2(t) \cdot m_2 + m_4. \end{aligned}$$
$$m_4(2) = 2m_4 + 6m_2^2; \quad m_4(3) = 3m_4 + 18m_2^2;$$

$$m_4(4) = 4m_4 + 36m_2^2; \quad m_4(5) = 5m_4 + 60m_2^2.$$
$$m_4(t) = t \cdot m_4 + 3t^{(2)}m_2^2.$$
$$m_5(t+1) = m_5(t) + 10m_3(t) \cdot m_2 + 10m_2(t) \cdot m_3 + m_5.$$
$$\begin{aligned} m_5(2) &= 2m_5 + 20m_2 \cdot m_3; \quad m_5(3) = 3m_5 + 60m_2 \cdot m_3; \\ m_5(4) &= 4m_5 + 120m_2 \cdot m_3; \quad m_5(5) = 5m_5 + 200m_2 \cdot m_3. \end{aligned}$$
$$m_5(t) = t \cdot m_5 + 10t^{(2)} m_3 \cdot m_2.$$
$$m_1(t) = 0$$

[illegible]

[illegible]

[illegible]

[illegible]

$$m_6(t) = t \cdot m_6 + 15t^{(2)}m_4 \cdot m_2 + 10t^{(2)}m_3^2 + 15t^{(3)}m_2^3 \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (v)$$

$$m_7(t) = t \cdot m_7 + 21t^{(2)}m_5 \cdot m_2 + 35t^{(2)}m_4 \cdot m_3 + 105t^{(3)}m_3 \cdot m_2^2 \quad . \quad . \quad . \quad . \quad . \quad . \quad (\text{vi})$$

$$m_8(t) = t \cdot m_8 + 28t^{(2)}m_6 \cdot m_2 + 56t^{(2)}m_5 \cdot m_3 + 35t^{(2)}m_4^2 \\ + 210t^{(3)}m_4 \cdot m_2^2 + 280t^{(3)}m_3^2 \cdot m_2 + 105t^{(4)}m_2^4 \quad (\text{vii})$$

$$E\left(\frac{X_a}{a}\right)^k = \frac{1}{a^k} E(X_a)^k.$$

The scalar factor  $a^k$  cancels out in the denominator and numerator of the Pearson  $\beta$ -coefficients. So Pearson coefficients of the same order are identical for score-sum and score-mean or proportionate-score distributions ; and we now have all the data for expressing Pearson coefficients





(a) *Poisson Universe*

That of the unit sample is a highly skew distribution which is also *leptokurtic* ( $\beta_2 > 3$ ), i.e. steeper in the region of the mode than the normal. For successive values in the range  $x = 0, 1, 2 \dots$ , the frequency equation is

$$y = \frac{e^{-M} M^x}{x!}.$$

The skewness depends on the single parameter in the above, i.e. the mean value ( $M$ ) of the score  $x$ . From the results already touched on in Chapter 6 of Vol. I, and set forth more fully in (x) of 14.03, we obtain the following values for the Pearson coefficients, when  $M = 1$ :

$$({}_1)\beta_1 = 1; \quad ({}_1)\beta_2 = 4; \quad ({}_1)\beta_3 = 11; \quad ({}_1)\beta_4 = 41.$$

By recourse to the preceding theorem, we obtain for the distribution of the score mean of 10-fold and 20-fold samples,

Sample size	$\beta_1$	$\beta_2$	$\beta_3$	$\beta_4$
10	0.10	3.10	1.01	17.51
20	0.05	3.05	0.5025	16.16
Normal	0.0	3.0	0.0	15.0

(b) *Rectangular Universe*

For an  $n$ -fold universe of score values  $1, 2, 3 \dots n$ , the expression for the score-frequency of the unit sample is

$$y = \frac{1}{n}.$$

The odd mean moments are all of zero value, as demonstrated on p. 594, and

$$({}_1)\beta_2 = \frac{9}{5} - \frac{12}{5(n^2 - 1)}; \quad ({}_1)\beta_4 = \frac{27}{7} - \frac{108}{7(n^2 - 1)} + \frac{144}{7(n^2 - 1)^2};$$

$$({}_1)\beta_6 = 9 - \frac{73}{n^2 - 1} + \frac{1296}{5(n^2 - 1)^2} - \frac{1728}{5(n^2 - 1)^3}.$$

For the 6-fold ( $n = 6$ ) rectangular universe of the ordinary cubical die we therefore have

$$({}_1)\beta_2 = 1.731; \quad ({}_1)\beta_4 = 3.422; \quad ({}_1)\beta_6 = 7.146.$$

For the mean score distribution of the  $t$ -fold toss, we obtain from the foregoing theorem,

No. of tosses	$\beta_1$	$\beta_2$	$\beta_3$	$\beta_4$	$\beta_5$	$\beta_6$
6	0	2.789	0	12.036	0	65.103
12	0	2.895	0	13.466	0	84.586
18	0	2.930	0	13.966	0	91.002
Normal	0	3.0	0	15.0	0	105.0

(c) *A Binomial Universe*

The determination of the higher moments of the Binomial distribution, though elementary, is somewhat laborious by more usual methods. The relations established above permit us to compute both moments and Pearson coefficients by a shorter route. The procedure depends on a theorem established in Chapter 7 of Vol. I, *viz.*: if successive terms of the binomial  $(p + q)^a$  define the *unit* sample score distribution of an  $(a + 1)$ -fold universe, successive terms of the

binomial  $(p + q)^{at}$  define the random distribution of the score-sum or mean score of the  $t$ -fold sample therefrom. If we make  $a = 1$ , denoting by  ${}_{(1)}m_k$  the  $k$ th mean moment of the distribution whose definitive binomial is  $(p + q)^1$ , the corresponding moment  ${}_{(t)}m_k$  is that of the distribution of the  $t$ -fold sample score in the taxonomic domain of a binary universe. For its unit sample the mean is then  $p$  and the distribution is simply

Raw Score ( $x$ ) . . . . .	0	1
Score Deviation . . . . .	$-p$	$(1 - p) = q$
Frequency ( $y$ ) . . . . .	$q$	$p$

Thus the  $k$ th mean moment of the unit sample distribution is

$$m_k = q(-p)^k + pq^k.$$

Whence we derive

$$m_2 = qp^2 + pq^2 = pq(p + q) = pq;$$

$$m_3 = pq^3 - qp^3 = pq(q^2 - p^2) = pq(q - p)(q + p) = pq(q - p);$$

$$m_4 = qp^4 + pq^4 = pq(p^3 + q^3) = pq(p^2 - pq + q^2) = pq(1 - 3pq);$$

$$m_5 = pq^5 - qp^5 = pq(q^4 - p^4) = pq(q^2 - p^2)(q^2 + p^2) = pq(q - p)(1 - 2pq);$$

$$m_6 = qp^6 + pq^6 = pq(1 - 5p^4q - 10p^3q^2 - 10p^2q^3 - 5pq^4) = pq(1 - 5pq + 5p^2q^2);$$

$$m_7 = pq^7 - qp^7 = pq(q^6 - p^6) = pq(q^3 - p^3)(q^3 + p^3) = pq(q - p)(1 - pq)(1 - 3pq);$$

$$m_8 = qp^8 + pq^8 = pq\{1 - 7pq(1 - 2pq + p^2q^2)\}.$$

We may write the general formulae for the Pearson coefficients as follows:

$${}_{(1)}\beta_{2k-1} = \frac{m_3 \cdot m_{2k+1}}{m_2^{k+2}} = \frac{(q - p)(q^{2k} - p^{2k})}{p^k q^k};$$

$${}_{(1)}\beta_{2k-2} = \frac{m_{2k}}{m_2^k} = \frac{p^{2k-1} + q^{2k-1}}{p^{k-1} q^{k-1}}.$$

Hence we get

$${}_{(1)}\beta_1 = \frac{(q - p)^2}{pq};$$

$${}_{(1)}\beta_2 = \frac{1 - 3pq}{pq};$$

$${}_{(1)}\beta_3 = \frac{(q - p)^2(1 - 2pq)}{p^2 q^2};$$

$${}_{(1)}\beta_4 = \frac{1 - 5pq + 5p^2 q^2}{p^2 q^2};$$

$${}_{(1)}\beta_5 = \frac{(q - p)^2(1 - pq)(1 - 3pq)}{p^3 q^3}; \quad {}_{(1)}\beta_6 = \frac{1 - 7pq + 14p^2 q^2 - 7p^3 q^3}{p^3 q^3}.$$

We thus obtain

$${}_{(t)}\beta_1 = \frac{(q - p)^2}{tpq};$$

$${}_{(t)}\beta_2 = 3 + \frac{1 - 6pq}{tpq};$$

$${}_{(t)}\beta_3 = \frac{\{1 + pq(10t - 12)\}(q - p)^2}{t^2 p^2 q^2};$$

$${}_{(t)}\beta_4 = 15 + \frac{1 - 30pq(1 - 4pq) + 5tpq(5 - 26pq)}{t^2 p^2 q^2};$$



$${}_{(t)}\beta_5 = \frac{(q-p)^2\{(1-pq)(1-3pq) + 7(t-1)pq(8-51pq+15tpq)\}}{t^3p^3q^3};$$

$${}_{(t)}\beta_6 = 105 + \frac{1-7pq(1-2pq+16p^2q^2)}{t^3p^3q^3} + \frac{7(t-1)}{t^3p^3q^3}[17-238pq+704p^2q^2+10tpq(7-34pq)].$$

When  $p = \frac{1}{2} = q$ ,

$${}_{(1)}\beta_1 = 0; \quad {}_{(1)}\beta_2 = 1;$$

$${}_{(1)}\beta_3 = 0; \quad {}_{(1)}\beta_4 = 1;$$

$${}_{(1)}\beta_5 = 0; \quad {}_{(1)}\beta_6 = 1.$$

In Chapter 3 of Vol. I we have seen that the normal is a good fitting curve at the  $2\sigma$  level for the discrete distributions whose definitive binomials are respectively

$$(a) \left(\frac{1}{2} + \frac{1}{2}\right)^{16}; \quad (b) \left(\frac{9}{10} + \frac{1}{10}\right)^{100}.$$

From the above values for the Pearson coefficients of the 2-class distribution defined by the unit binomial  $\left(\frac{1}{2} + \frac{1}{2}\right)^1$ , we obtain the following results for the distributions of unit and 10-fold sample mean scores, when the binomial definitive of a universe of 17 score classes is  $\left(\frac{1}{2} + \frac{1}{2}\right)^{16}$ .

Sample size	$\beta_1$	$\beta_2$	$\beta_3$	$\beta_4$
1	0	2.875	0	13.2
10	0	2.988	0	14.813

For a universe of 101 score classes specified by successive terms of the binomial  $\left(\frac{9}{10} + \frac{1}{10}\right)^{100}$ , we obtain

	$\beta_1$	$\beta_2$	$\beta_3$	$\beta_4$
1	0.07	3.05	0.71	16.469
10	0.007	3.005	0.071	15.148

It is evident from these results that the distribution of the mean of a sample as small as 10 closely conforms to the normal pattern for a discrete distribution as flat as may be, and for a relatively steep and skew unimodal distribution. It is also noteworthy that the range of the unit sampling distribution consistent with this assertion may be very restricted. From that viewpoint, the following situation is instructive.

#### (d) The "Burette" Universe

The term "burette" universe here signifies one of a type of situations which not uncommonly arise in the laboratory, when repeated estimations ring the changes on only 3 consecutive scale divisions consonant with competent workmanship. We thus suppose that the unit sampling distribution involves only 3 score values, which we may label  $-1$ ,  $0$  and  $+1$ , if the scores run consecutively. We shall first suppose that we obtain the central score twice as often as otherwise so that the specification of the symmetrical unit sample distribution is

Score	.	.	.	$-1$	$0$	$+1$
Frequency	.	.	.	$\frac{1}{6}$	$\frac{2}{3}$	$\frac{1}{6}$

Evidently all Pearson coefficients of *odd* order have zero value. As an exercise the reader may check the following :

$${}_{(1)}\beta_2 = 3; \quad {}_{(1)}\beta_4 = 9; \quad {}_{(1)}\beta_6 = 27.$$

## 3-FOLD SAMPLE DISTRIBUTION FROM A SYMMETRICAL BURETTE UNIVERSE

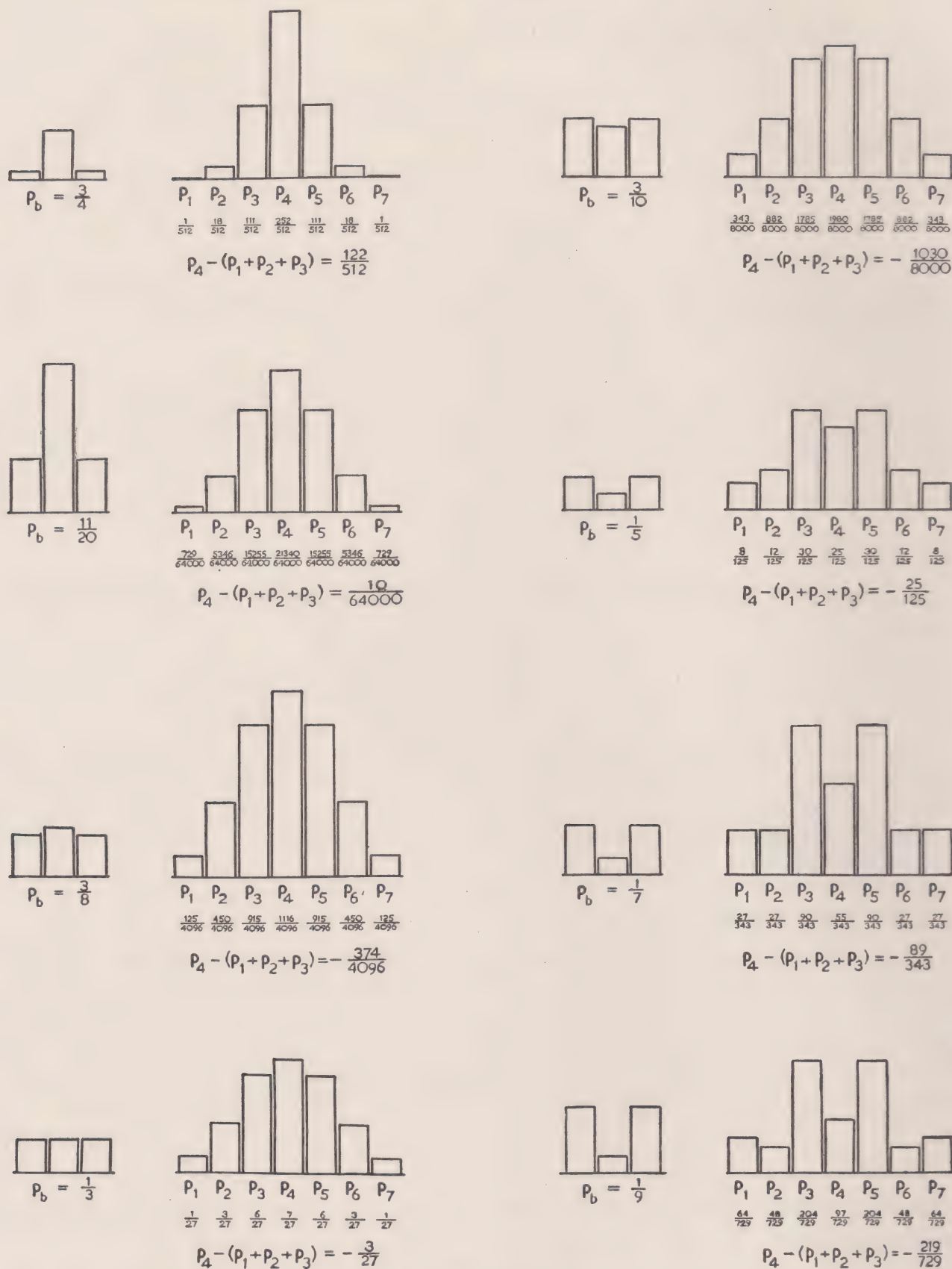


FIG. 107. 2-Fold Sample Distributions from Symmetrical Burette Universes.



2 - FOLD SAMPLE DISTRIBUTION FROM A SYMMETRICAL BURETTE UNIVERSE

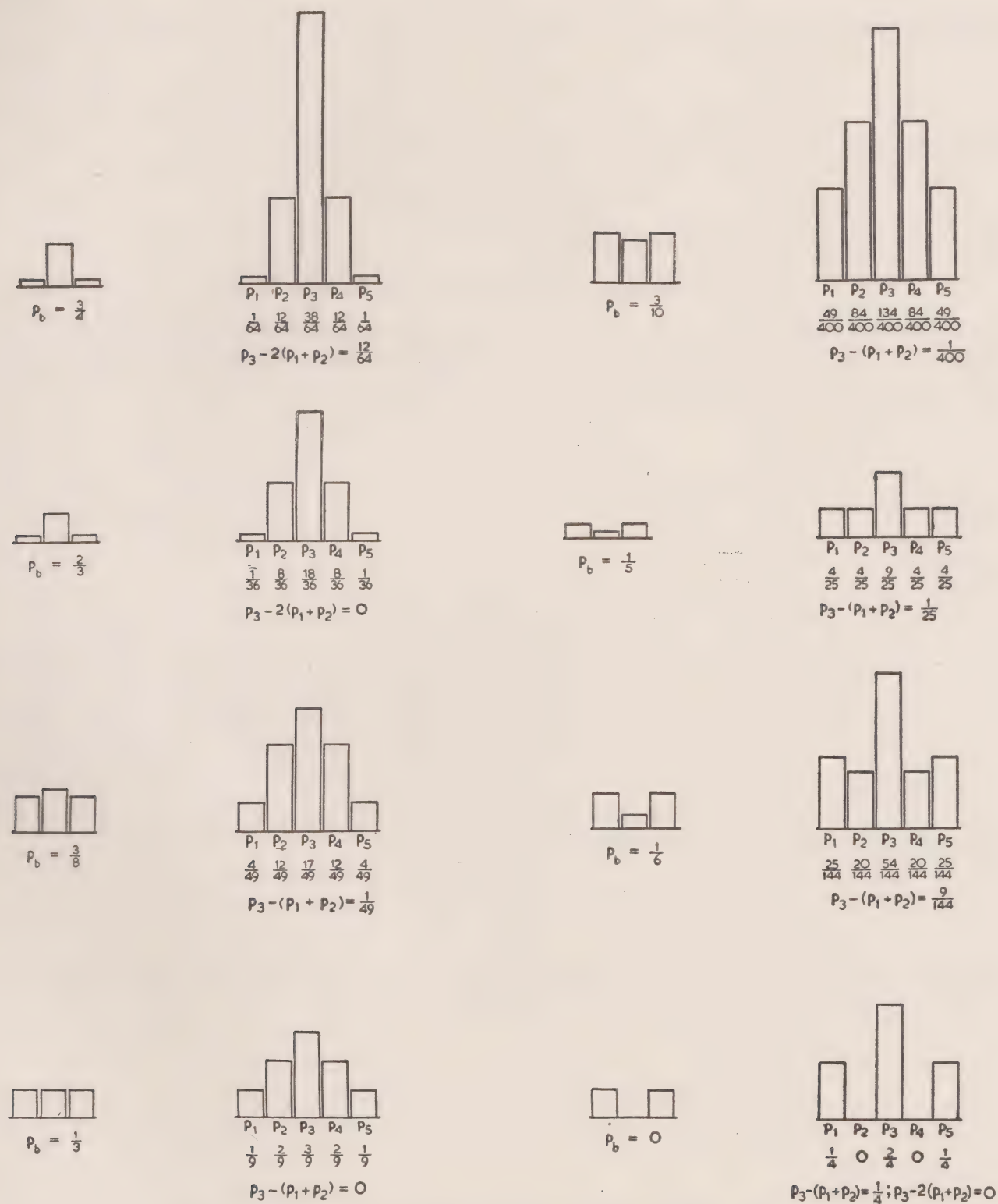


FIG. 108. 3-Fold Sample Distributions from Symmetrical Burette Universes.

Whence we obtain for the 10-fold sample, the following remarkable correspondence with the normal values :

$${}_{(10)}\beta_2 = 3; \quad {}_{(10)}\beta_4 = 14.94; \quad {}_{(10)}\beta_6 = 103.41.$$

For the 3-class universe with unit scale we may write the u.s.d. more generally as below :

$$\begin{array}{ccccccc} \text{Score} & . & . & . & -1 & 0 & +1 \\ \text{Frequency} & . & . & . & p_a & (1 - p_a - p_c) & p_c \end{array}$$

If the distribution is symmetrical ( $p_a = p_c$ ) the mean is zero, and

$$m_{2k+1} = 0; \quad m_{2k+2} = 2p_a.$$

Thus odd order Pearson coefficients vanish and

$$\beta_2 = (2p_a)^{-1}; \quad \beta_4 = (2p_a)^{-2}; \quad \beta_6 = (2p_a)^{-3}.$$

For symmetrical distributions we thus derive results shown in Table 3 which brings into focus the issue dealt with in 14.07 : how nearly must the numerical values of the coefficients correspond with those of the normal to guarantee a good normal fit ?

TABLE 3

Definitive Trinomial	$(\frac{2}{5} + \frac{1}{5} + \frac{2}{5})^t$			$(\frac{1}{3} + \frac{1}{3} + \frac{1}{3})^t$			$(\frac{1}{4} + \frac{1}{2} + \frac{1}{4})^t$			$(\frac{1}{6} + \frac{4}{6} + \frac{1}{6})^t$			$(\frac{1}{10} + \frac{8}{10} + \frac{1}{10})^t$		
Sample Size (t)	$\beta_2$	$\beta_4$	$\beta_6$	$\beta_2$	$\beta_4$	$\beta_6$	$\beta_2$	$\beta_4$	$\beta_6$	$\beta_2$	$\beta_4$	$\beta_6$	$\beta_2$	$\beta_4$	$\beta_6$
Unit	1.25	1.56	1.95	1.5	2.25	3.38	2.0	4.0	8.0	3	9.0	27.0	5.0	25.0	125.0
2	2.13	5.08	12.54	2.25	6.19	18.14	2.5	8.5	39.5	3	13.5	74.25	4.0	25.0	212.5
4	2.56	9.24	39.09	2.63	9.93	46.07	2.75	10.4	54.2	3	14.3	95.91	3.5	21.25	184.06
6	2.71	10.98	55.76	2.75	11.52	61.61	2.87	12.61	75.2	3	14.83	100.75	3.33	19.4	163.43
10	2.83	12.50	72.71	2.85	12.85	76.88	2.9	13.54	85.94	3	14.94	103.41	3.25	17.8	142.82
20	2.91	13.79	87.76	2.93	13.90	90.11	2.95	14.26	95.00	3	15.0	104.59	3.1	16.45	124.95
40	2.96	14.35	96.10	2.96	14.44	98.14	2.98	14.99	99.88	3	15.0	104.90	3.05	15.74	115.24

When the distribution is not symmetrical, we may write the moments about zero in the following form by means of the substitution  $p_c = hp_a$  :

$$\begin{aligned} \mu_{2k+1} &= p_c(1)^{2k+1} + p_a(-1)^{2k+1} = p_a(h-1); \\ \mu_{2k+2} &= p_c(1)^{2k+2} + p_a(-1)^{2k+2} = p_a(h+1). \end{aligned}$$

In virtue of the identity of odd zero moments and of even zero moments, the expressions for mean moments involve only  $\mu_1 = p_a(h-1)$  and  $\mu_2 = p_a(h+1)$ , thus :

$$\begin{aligned} m_2 &= \mu_2 - \mu_1^2; \\ m_3 &= \mu_3 - 3\mu_2\mu_1 + 2\mu_1^3 = \mu_1(1 - 3\mu_2 + 2\mu_1^2); \\ m_4 &= \mu_4 - 4\mu_3\mu_1 + 6\mu_2\mu_1^2 - 3\mu_1^4 = \mu_2(1 + 6\mu_1^2) - \mu_1^2(4 + 3\mu_1^2). \end{aligned}$$

Thus we derive

$$\beta_2 = \frac{(h+1)[1 + 6p_a^2(h-1)^2] - p_a(h-1)^2[4 + 3p_a^2(h-1)^2]}{p_a(h+1)^2 - 2p_a^2(h^2-1)(h-1) + p_a^3(h-1)^4}.$$

The reader may derive other coefficients as an exercise.



EXERCISE 14.05

1. If  $\mu_k$  is the  $k$ th zero moment of the unit sample distribution and  $\mu_k(r)$  is that of the  $r$ -fold sample distribution with replacement, show that

$$\begin{aligned}\mu_2(r) &= r\mu_2 + r^{(2)}\mu_1^2; \\ \mu_3(r) &= r\mu_3 + 3r^{(2)}\mu_1\mu_2 + r^{(3)}\mu_1^3; \\ \mu_4(r) &= r\mu_4 + 4r^{(2)}\mu_1\mu_3 + 3r^{(2)}\mu_2^2 + 6r^{(3)}\mu_1^2\mu_2 + r^{(4)}\mu_1^4.\end{aligned}$$

2. Use the results of 1 to show that the first four zero moments of the distribution whose definitive binomial is  $(q + p)^r$  are

$$\begin{aligned}\mu_1 &= rp; \quad \mu_2 = rp + r^{(2)}p^2; \\ \mu_3 &= 2p + 3r^{(2)}p^2 + r^{(3)}p^3; \quad \mu_4 = rp + 7r^{(2)}p^2 + 6r^{(3)}p^3 + r^{(4)}p^4.\end{aligned}$$

3. Find the first eight mean moments of the raw-score distribution whose definitive binomial is  $(q + p)^r$ .

4. For the 3-class distribution of scores  $-1, 0$  and  $+1$  with frequencies  $p_a, p_b$  and  $p_c$  give the values of  $\beta_1, \beta_2, \beta_3$  and  $\beta_4$  for the  $r$ -fold sample mean-score distribution.

5. Tabulate the values of  $\beta_1, \beta_2, \beta_3$  and  $\beta_4$  for the distribution of Exercise 4 when  $r = 5$  and  $r = 10$  for  $p_a = \frac{1}{3}$  and  $p_b = \frac{1}{6}$ .

14.06 MOMENTS OF A DIFFERENCE DISTRIBUTION

In a preliminary way, we have examined in Chapter 6 of Vol I what moment-fitting curve is appropriate to describe the distribution of the difference between 2 independent binomial variates. In that context our concern was with only the first two Pearson coefficients. We shall now take up the same issue from the more general viewpoint of 11.06.

Before doing so, it is appropriate to clarify an issue we have not as yet dealt with. When our concern is to explore the null hypothesis that 2 samples come from one and the same binary universe, two courses are open in accordance with the schema below :

	Sample A	Sample B	Total
No. of Successes	$x_a$	$x_b$	$x_a + x_b$
No. of Failures	$a - x_a$	$b - x_b$	$N - x_a - x_b$
Total	$a$	$b$	$N$

In this set-up, our estimate of the proportion of successes in the putative common universe of the null hypothesis is

$$p_o = \frac{x_a + x_b}{N}.$$

The estimated mean numbers of successes for the two samples are therefore

$$M_a = a \cdot p_o \quad \text{and} \quad M_b = b \cdot p_o.$$

Accordingly, the estimated mean value of the raw-score difference  $(x_a - x_b) = D$  is

$$M_D = (a - b)p_o.$$

The estimated variances for samples of large size are

$$V_a = a \cdot p_o(1 - p_o) \quad \text{and} \quad V_b = b \cdot p_o(1 - p_o).$$

Accordingly, the estimated variance of the difference  $(x_a - x_b) = D$  is

$$V_D = (a + b)p_o(1 - p_o) = N \cdot p_o(1 - p_o).$$

Whence the square *standard* raw-score difference is

$$c_r^2 = \frac{(D - \overline{a - b \cdot p_o})^2}{N \cdot p_o(1 - p_o)} = \frac{4(b \cdot x_a - a \cdot x_b)^2}{N(x_a + x_b)(N - x_a - x_b)} \quad \dots \quad (i)$$

For the corresponding proportionate score difference we may write

$$d = \frac{x_a}{a} - \frac{x_b}{b} = \frac{b \cdot x_a - a \cdot x_b}{ab};$$

$$V_d = \frac{p_o(1 - p_o)}{a} + \frac{p_o(1 - p_o)}{b} = \frac{N \cdot p_o(1 - p_o)}{ab}.$$

Whence we may write for the square standard proportionate score difference

$$c_p^2 = \frac{N(b \cdot x_a - a \cdot x_b)^2}{ab(x_a + x_b)(N - x_a - x_b)} \quad \dots \quad (ii)$$

Accordingly, and if  $a = Kb$ , we derive

$$\frac{c_p^2}{c_r^2} = \frac{N^2}{4ab} = \frac{(K + 1)^2}{4K} \quad \dots \quad (iii)$$

The expression on the right is a minimum when  $K = 1$ , i.e. when the samples are equal, in which case  $c_r^2 = c_p^2$ . Otherwise,  $c_p^2 > c_r^2$ . If therefore the distribution of the score difference is approximately normal in either case, the proportionate score difference will give the higher assessment of odds against the null hypothesis unless  $a = b$ .

In this context it is not inappropriate to refer to a common misconception. Many statistical textbooks advocate the so-called *Chi-Square test* for assessing the credentials of the null hypothesis under discussion. The fact is that the assessment of odds w.r.t. the proportionate-score difference as prescribed above is exactly the same as by the Chi-Square test for 1 d.f., Chi-Square in this context being the square standard score. For the performance of the latter test,  $x$  being the cell score, we define Chi-Square ( $C$ ) in terms of the cell score  $x$  and the *observed* mean ( $M_x$ ) of its four values  $x_a$ ,  $(a - x_a)$ ,  $x_b$ ,  $(b - x_b)$  as

$$C = \sum \frac{(x - M_x)^2}{M_x}.$$

Thus we write

$$C = \frac{\left[ x_a - \frac{a(x_a + x_b)}{N} \right]^2}{\frac{a(x_a + x_b)}{N}} + \frac{\left[ x_b - \frac{b(x_a + x_b)}{N} \right]^2}{\frac{b(x_a + x_b)}{N}}$$

$$+ \frac{\left[ (a - x_a) - \frac{a(N - x_a - x_b)}{N} \right]^2}{\frac{a(N - x_a - x_b)}{N}} + \frac{\left[ (b - x_b) - \frac{b(N - x_a - x_b)}{N} \right]^2}{\frac{b(N - x_a - x_b)}{N}}$$



$$\begin{aligned}
&= \frac{(bx_a - ax_b)^2}{aN(x_a + x_b)} + \frac{(bx_a - ax_b)^2}{bN(x_a + x_b)} + \frac{(bx_a - ax_b)^2}{aN(N - x_a - x_b)} + \frac{(bx_a - ax_b)^2}{bN(N - x_a - x_b)} \\
&= \frac{(bx_a - ax_b)^2}{N} \left[ \frac{(a+b)}{ab(x_a + x_b)} + \frac{(a+b)}{ab(N - x_a - x_b)} \right] \\
&= \frac{N(bx_a - ax_b)^2}{ab(x_a + x_b)(N - x_a - x_b)}.
\end{aligned}$$

The last expression is identical with (ii); and it is obviously immaterial whether we care to consult the table of Chi-Square for 1 d.f. in order to assess the significance of the square standard score or that of the normal integral to assess that of the standard score itself. The two tests must necessarily give the same result; but will not give the same result as the normal test for the raw-score difference. The procedure last mentioned will in fact give a more conservative estimate of the odds against the null hypothesis unless the size of the 2 samples is the same.

In Chapter 4 of Vol. I we have seen reason to suppose that the raw-score difference distribution tends to normality for small values of  $r_a$  and  $r_b$  more closely than that of the proportionate-score difference, which has peculiar features for co-prime samples (Chapter 4, Vol. I). Consequently, the relative advantage of a test based on the latter procedure is not clear-cut without further investigation of the approach of the two distributions to the normal. We may explore this by the method of 11.06 without confining our attention to the binomial case.

We first recall the fact that the difference between a score difference and its mean value is the difference between the two score deviations, i.e.

$$\begin{aligned}
E(x_a - x_b) &= E(x_a) - E(x_b), \\
\therefore (x_a - x_b) - E(x_a - x_b) &= [x_a - E(x_a)] - [x_b - E(x_b)], \\
\therefore (x_a - x_b) - E(x_a - x_b) &= X_a - X_b.
\end{aligned}$$

We therefore write the  $k$ th mean moments of the raw-score and proportionate-score difference distributions as below:

*Raw-score difference:*

$$m_k(D) = E(X_a - X_b)^k \quad \dots \quad (iv)$$

*Proportionate-score difference:*

$$m_k(d) = E\left(\frac{X_a}{a} - \frac{X_b}{b}\right)^k \quad \dots \quad (v)$$

As in 13.05 we shall write the  $k$ th mean moment of the score-sum distribution of the  $t$ -fold sample as  $m_k(t)$ , and that of the unit sample distribution as  $m_k$  using the results of 14.05 to express the former in terms of the latter. For the raw-score difference distribution we then have

$$\begin{aligned}
m_2(D) &= m_2(a) + m_2(b) = (a + b)m_2; \\
m_3(D) &= m_3(a) - m_3(b) = (a - b)m_3; \\
m_4(D) &= m_4(a) + 6m_2(a) \cdot m_2(b) + m_4(b) \\
&= (a + b)m_4 + 3[a^{(2)} + 2ab + b^{(2)}]m_2^2 \\
&= (a + b)m_4 + 3(a + b)^{(2)}m_2^2.
\end{aligned}$$





$$\frac{\beta_2(d) - 3}{\beta_2(S) - 3} = \frac{K^2 - K + 1}{K} = (K - 1) + \frac{1}{K} \quad \text{(xii)}$$

From (ix) we see that  $\beta_1(d)$  must be zero as for the normal distribution if the samples are of equal size ( $a = b$  and  $K = 1$ ); but it is not otherwise necessarily less than  $\beta_1(S)$ . The expression on the right of (xi) exceeds unity when  $K \simeq 2.6$ . This means that the proportionate-score difference distribution w.r.t.  $a$ -fold and  $b$ -fold samples will be *less skew* than the score-sum distribution of the  $(a + b)$ -fold sample, if the ratio of the two samples does not appreciably exceed 2.6. Otherwise, it is more skew. The expression on the right of (xii) always exceeds unity, if  $K \geq 1$ , so that the divergence of  $\beta_2(d)$  from the normal value will always be greater than that of  $\beta_2(S)$  and hence greater than that of  $\beta_2(D)$ . By comparison of (ix) and (vi) we see that

$$\frac{\beta_1(d)}{\beta_1(D)} = \frac{(K + 1)^2}{K} = K + 2 + \frac{1}{K}.$$

Unless  $K = 1$  (when both distributions are symmetrical), this means that the proportionate-score difference distribution is more skew than that of the raw-score. It is evident that  $\beta_1$  and  $\beta_2$  for both distributions rapidly approach the normal values of 0 and 3 as  $b$ , the size of the smaller sample, becomes large. For we may write them in the form :

$$\begin{aligned} \beta_1(D) &= \frac{(K - 1)^2}{(K + 1)^3 \cdot b} \cdot \beta_1 \quad \text{and} \quad \beta_1(d) = \frac{(K - 1)^2}{K(K + 1) \cdot b} \cdot \beta_1; \\ \beta_2(D) - 3 &= \frac{\beta_2 - 3}{(K + 1) \cdot b} \quad \text{and} \quad \beta_2(d) - 3 = \frac{(K^2 - K + 1)}{K(K + 1) \cdot b} (\beta_2 - 3). \end{aligned}$$

We may investigate higher Pearson coefficients of the two difference distributions in the same way, and obtain

$$\beta_3(D) = \frac{(a - b)^2}{(a + b)^2} \beta_3(S); \quad \beta_5(D) = \frac{(a - b)^2}{(a + b)^2} \beta_5(S) \quad \text{(xiii)}$$

$$\beta_4(D) = \beta_4(S) - \frac{40ab}{(a + b)^2} \beta_1(S) \quad \text{(xiv)}$$

$$\beta_6(D) = \beta_6(S) - \frac{224ab}{(a + b)^4} \{\beta_3 + 5(a + b - 2)\beta_1\} \quad \text{(xv)}$$

Thus  $\beta_4(D) = \beta_4(S)$  and  $\beta_6(D) = \beta_6(S)$  if the unit sample distribution is symmetrical. We also see that the values of  $\beta_3$  and  $\beta_5$  for the raw-score difference distribution will always be less than those of the corresponding coefficients of the score-sum distribution unless the unit sample distribution is symmetrical, in which case all coefficients of odd order vanish in both cases. If the unit sample distribution is skew and leptokurtic the values of  $\beta_4$  and  $\beta_5$  for the raw-score distribution will be smaller than those for that of the score-sum, i.e. the raw-score distribution will be less leptokurtic as well as less skew. Our investigation of the moments of the raw-score difference distribution thus leads us to the conclusion that it approaches more closely to the normal than does the score-sum distribution for the  $(a + b)$ -fold sample; but we have already seen that no comparable straightforward statement applies to the difference distribution of the proportionate score.

Expressions for higher Pearson coefficients of the latter are unwieldy and the derivation is laborious. Accordingly, we shall cite only approximate expressions, neglecting terms which are trivial. When  $(a + b)$  is large :

$$\begin{aligned}\beta_3(d) &\simeq \frac{10(K-1)^2}{K(K+1)b}\beta_1; & \beta_5(d) &\simeq \frac{105(K-1)^2}{K(K+1)b}\beta_1; \\ \beta_4(d) &\simeq 15 + \frac{15(K^2 - K + 1)}{K(K+1)b}(\beta_2 - 3) + \frac{10(K-1)^2}{K(K+1)b}\beta_1; \\ \beta_6(d) &\simeq 105 + \frac{210K(\beta_2 - 3)}{(K+1)^3 \cdot b} + \frac{280(K^4 + 1)}{K(K+1)^3 \cdot b}\beta_1.\end{aligned}$$

Evidently, all the above approach the normal as limiting values when  $b$  (and hence also  $a$ ) is large.

#### 14.07 NORMAL APPROXIMATIONS

How nearly the standardised moments of a discrete distribution approach those of the normal is a purely algebraic issue, if the algebraic pattern of the former is specifiable. How close the correspondence must be to justify recourse to the normal as a descriptive function is largely an empirical one ; but it may be possible to limit the ground for numerical exploration by reference to an alternative standard. Many theoretical distributions whose moments are specifiable (e.g. the *Poisson*) closely approach the normal with suitable choice of parameters ; and if they are amenable to tabulation it is a simple matter to choose one as a *yardstick distribution*. For instance, we may easily derive from the tables of the Poisson function what value of its single parameter  $M$  ensures a percentage error no greater than  $e$  for summation of all values up to the  $2\sigma$  level. We may then ask what conditions ensure that the standardised moments of the binomial variate defined by successive terms of  $(q + p)^r$  lie closer to the normal than do those of the Poisson yardstick distribution.

When we speak of a binomial variate so defined in the most general sense, i.e. without restriction on the origin or scale, we assume a set of scores of range  $a$  to  $a + r\Delta x$  so that  $r_{(x)} \cdot p^x q^{r-x}$  is the frequency of the score  $a + x\Delta x$ . If we are speaking of the raw-score of an  $r$ -fold sample from an infinite 2-class universe this means that  $a = 0$  and  $\Delta x = 1$ . If we are speaking of the proportionate-score deviation of such a sample  $a = -p$  and  $\Delta x = r^{-1}$ . If we are speaking of the mean score of the 3-fold toss of a tetrahedral die with face pips 2, 4, 4, 6 we have  $a = 2$  and  $\Delta x = \frac{2}{3}$ . Here it suffices to consider the situation which arises when  $a = 0$  and  $\Delta x = 1$  since  $a$  does not affect the value of the mean moments and the appropriate power of  $\Delta x$  appears as a scalar factor common to both the numerator and the denominator of the Pearson coefficients. Since also  $(q + p)^{ra}$  defines the sampling distribution of the  $ra$ -fold score-sum from an infinite two class universe and that of the  $r$ -fold score-sum from a universe of  $(a + 1)$  classes whose frequencies tally with successive terms of  $(q + p)^a$ , it will suffice to define the moments of any binomial universe in terms of those of the universe of 2 classes.

With that end in view it will be convenient to write  $q = mp$ , so that  $p = (m + 1)^{-1}$ , and if  $rp = M$ ,  $r = M(m + 1)$ . If  $q + p$  defines the u.s.d. of score values 0, 1 and deviations  $-p$ ,  $q = (1 - p)$ , we may write :

$$\begin{aligned}m_{2k+1} &= q(-p)^{2k+1} + pq^{2k+1} = pq(q^{2k} - p^{2k}) = \frac{m(m^{2k} - 1)}{(m + 1)^{2k+2}}; \\ m_{2k} &= q(-p)^{2k} + pq^{2k} = pq(q^{2k-1} + p^{2k-1}) = \frac{m(m^{2k-1} + 1)}{(m + 1)^{2k+1}}.\end{aligned}$$



Whence

$$\begin{aligned}\beta_1 &= \frac{(m-1)^2}{m}; \quad \beta_2 = \frac{m^2 - m + 1}{m}; \\ \beta_3 &= \frac{(m-1)^2(m^2 + 1)}{m^2}; \quad \beta_4 = \frac{m^4 - m^3 + m^2 - m + 1}{m^2}; \\ \beta_5 &= \frac{(m-1)^2(m^4 + m^2 + 1)}{m^3}; \quad \beta_6 = \frac{m^6 - m^5 + m^4 - m^3 + m^2 - m + 1}{m^3}.\end{aligned}$$

From the above we see that odd Pearson coefficients vanish only if  $m = 1$  ( $p = \frac{1}{2} = q$ ) as we know; and we can define the value of  $m$  which confers normal kurtosis by putting

$$\beta_2 = 3 = \frac{m^2 - m + 1}{m} \quad \text{so that} \quad m^2 - 4m + 1 = 0.$$

The roots of the above are approximately 3.73 and 0.27 corresponding to  $p \simeq 0.21$  and 0.79 within which range  $\beta_2 < 3$  and the distribution is platykurtic. Outside this range the distribution is leptokurtic and skew as is true of the Poisson. For the distribution defined by successive terms of  $(q + p)^r$  in terms of  $m$  and  $M = rp$  we have the following values of the Pearson coefficients tending to the Poisson limits and lying consistently *between* those of the Poisson distribution and those of the normal outside the range  $m > 3.73$ :

$$\begin{aligned}{}_r\beta_1 &= \frac{(m-1)^2}{Mm(m+1)} \simeq \frac{1}{M} \quad \text{in the limit}; \\ {}_r\beta_2 &= 3 + \frac{1}{M} - \frac{5m-1}{Mm(m+1)} \simeq 3 + \frac{1}{M} \quad \text{in the limit}; \\ {}_r\beta_3 &= \frac{(m-1)^2(m^2 - 10m + 1)}{M^2m^2(m+1)^2} + \frac{10(m-1)^2}{Mm(m+1)} \simeq \frac{1}{M^2} + \frac{10}{M} \quad \text{in the limit}; \\ {}_r\beta_4 &= 15 + \frac{25(m^2 - m + 1)}{Mm(m+1)} + \frac{(m^2 - m + 1)^2}{M^2m^2(m+1)^2} - \frac{24(m^2 - m + 1)}{M^2m(m+1)^2} - \frac{55}{M(m+1)} \\ &\quad + \frac{39}{M^2(m+1)^2} \simeq 15 + \frac{25}{M} + \frac{1}{M^2} \quad \text{in the limit}.\end{aligned}$$

Expressions for Pearson coefficients of higher orders may be obtained in a similar form, but are very unwieldy. We here cite the Poisson limiting forms:

$${}_r\beta_5 \simeq \frac{105}{M} + \frac{56}{M^2} + \frac{1}{M^3}; \quad {}_r\beta_6 \simeq 105 + \frac{490}{M} + \frac{119}{M^2} + \frac{1}{M^3}.$$

The above relations presuppose  $m > 1$ , so that  $q > p$  as is true of the Poisson distribution. Now the histogram of the distribution whose definitive binomial is  $(q + p)^r$  is the mirror image of that of the distribution whose definitive binomial is  $(p + q)^r$ . For every value  $m = k$  in the range  $m > 3.73$ , there will thus be in the range  $m < 0.27$  a value  $m = k^{-1}$  definitive of a distribution with identical Pearson coefficients. The above relations thus hold good for the range  $m < 0.27$  if we reverse the score order, i.e. we put  $M = rq$  when  $q < p$ . For a given value of  $M = rp$  when  $p < q$  (or  $M = rq$  when  $q < p$ ), we may therefore say that the variate

whose definitive binomial is  $(q + p)^r$  has mean moments nearer to their normal values than those of the Poisson distribution with the same parameter  $M$  if  $p$  lies outside the range stated. Subject to this restriction we may therefore say that the normal is a satisfactory fitting curve for a binomial variate if  $M$  exceeds the value for which the Poisson distribution tallies as closely with the normal as a satisfactory fit implies in the context. Table 4 shows how closely the Poisson distribution does in fact correspond to that of the normal at prescribed significance levels for three different values of  $M$ . Tables 5 and 6 explore the intermediate zone of  $p$  values consistent with platykurtosis of a binomial variate. Table 7 shows values of the Pearson coefficients for binomial variates with assigned values of  $M$ . In Tables 4-6, as in other tables of this chapter, the column headed *exact* under frequency refers to that of the discrete distribution, and the column marked *normal* next to it cites the corresponding ordinate of the normal curve. Under *cumulative frequency*, the column headed *normal* refers to areas bounded by the corresponding ordinate on either side of the mean after making the appropriate half interval correction.

TABLE 4

*Comparison of Cumulative Frequencies of Normal and Poisson ( $M = \sigma^2$ ) Distributions*

Raw-Score Deviation $X$	$M = 6 ; \beta_1 = 0.1\dot{6} ; \beta_2 = 3.1\dot{6}.$			$M = 10 ; \beta_1 = 0.10 ; \beta_2 = 3.10.$			$M = 15 ; \beta_1 = 0.0\dot{6} ; \beta_2 = 3.0\dot{6}.$		
	$\frac{X}{\sigma}$	Poisson	Normal	$\frac{X}{\sigma}$	Poisson	Normal	$\frac{X}{\sigma}$	Poisson	Normal
-12	.....	.....	.....	.....	.....	.....	-3.0984	0.0002	0.0015
-11	.....	.....	.....	.....	.....	.....	-2.8402	0.0008	0.0034
-10	.....	.....	.....	.....	.....	.....	-2.5820	0.0027	0.0068
-9	.....	.....	.....	-2.8460	0.0005	0.0036	-2.3238	0.0075	0.0141
-8	.....	.....	.....	-2.5298	0.0028	0.0089	-2.0656	0.0179	0.0264
-7	.....	.....	.....	-2.2136	0.0104	0.0199	-1.8074	0.0373	0.0467
-6	-2.4495	0.0025	0.0124	-1.8974	0.0293	0.0410	-1.5492	0.0697	0.0778
-5	-2.0412	0.0174	0.0310	-1.5811	0.0671	0.0774	-1.2910	0.1183	0.1227
-4	-1.6630	0.0620	0.0765	-1.2649	0.1302	0.1342	-1.0328	0.1846	0.1831
-3	-1.2247	0.1512	0.1537	-0.9487	0.2203	0.2146	-0.7746	0.2675	0.2593
-2	-0.8167	0.2851	0.2702	-0.6325	0.3329	0.3176	-0.5164	0.3631	0.3493
-1	-0.4082	0.4457	0.4191	-0.3162	0.4580	0.4382	-0.2582	0.4655	0.4486
0	0.0000	0.6063	0.5809	0.0000	0.5831	0.5618	0.0000	0.5679	0.5514
1	0.4082	0.7440	0.7298	0.3162	0.6968	0.6824	0.2582	0.6639	0.6507
2	0.8165	0.8473	0.8463	0.6325	0.7916	0.7854	0.5164	0.7486	0.7487
3	1.2247	0.9161	0.9235	0.9487	0.8645	0.8658	0.7746	0.8192	0.8169
4	1.6630	0.9574	0.9690	1.2649	0.9166	0.9226	1.0328	0.8749	0.8773
5	2.0412	0.9799	0.9876	1.5811	0.9513	0.9590	1.2910	0.9167	0.9222
6	2.4495	0.9912	0.9960	1.8974	0.9730	0.9801	1.5492	0.9466	0.9533
7	2.8577	0.9964	0.9989	2.2136	0.9858	0.9911	1.8074	0.9670	0.9736
8	3.2660	0.9986	0.9997	2.5298	0.9929	0.9964	2.0656	0.9803	0.9859
9	3.6743	0.9995	0.9999	2.8460	0.9966	0.9987	2.3238	0.9886	0.9932
10	4.0824	0.9998	.....	3.1623	0.9985	0.9996	2.5820	0.9936	0.9966
11	4.4906	0.9999	.....	3.4785	0.9994	0.9998	2.8402	0.9965	0.9985
12	.....	.....	.....	3.7947	0.9998	0.9999	3.0984	0.9981	0.9994
13	.....	.....	.....	4.1109	0.9999	.....	3.3566	0.9990	0.9997
14	.....	.....	.....	.....	.....	.....	3.6168	0.9994	0.9999
15	.....	.....	.....	.....	.....	.....	3.8730	0.9996	.....
16	.....	.....	.....	.....	.....	.....	4.1312	0.9997	.....



TABLE 5

 $(\frac{1}{2} + \frac{1}{2})^{20}$ ;  $M = 10$ ;  $\sigma = 2.2361$ ;  $\beta_1 = 0$ ;  $\beta_2 = 2.90$ 

Raw-Score Deviation $X$	$\frac{X}{\sigma}$	Frequency		Cumulative Frequency	
		Binomial	Normal	Binomial	Normal
0	0	0.1762	0.1784	0.1762	0.1770
$\pm 1$	$\pm 0.4472$	0.1602	0.1614	0.4966	0.4977
$\pm 2$	$\pm 0.8944$	0.1201	0.1194	0.7368	0.7344
$\pm 3$	$\pm 1.3416$	0.0739	0.0725	0.8846	0.8824
$\pm 4$	$\pm 1.7888$	0.0370	0.0361	0.9586	0.9558
$\pm 5$	$\pm 2.2360$	0.0148	0.0151	0.9882	0.9861
$\pm 6$	$\pm 2.6832$	0.0046	0.0049	0.9974	0.9963
$\pm 7$	$\pm 3.1304$	0.0011	0.0013	0.9996	0.9992
$\pm 8$	$\pm 3.5776$	0.0002	0.0003	1.0000	0.9999

TABLE 6

 $(\frac{3}{4} + \frac{1}{4})^{40}$ ;  $M = 10$ ;  $\sigma = 2.7386$ ;  $\beta_1 = 0.03$ ;  $\beta_2 = 2.983$ .

Raw-Score Deviation $X$	$\frac{X}{\sigma}$	Frequency		Cumulative Frequency	
		Binomial	Normal	Binomial	Normal
- 10	- 3.6510	0.0000	0.0002	0.0000	0.0003
- 9	- 3.2864	0.0001	0.0007	0.0001	0.0010
- 8	- 2.9212	0.0009	0.0021	0.0010	0.0021
- 7	- 2.5561	0.0037	0.0056	0.0047	0.0089
- 6	- 2.1909	0.0113	0.0133	0.0160	0.0223
- 5	- 1.8258	0.0273	0.0277	0.0433	0.0502
- 4	- 1.4606	0.0530	0.0499	0.0963	0.1006
- 3	- 1.0955	0.0857	0.0799	0.1820	0.1806
- 2	- 0.7303	0.1179	0.1116	0.2999	0.2919
- 1	- 0.3651	0.1398	0.1363	0.4397	0.4275
0	0.0000	0.1444	0.1457	0.5841	0.5725
1	0.3657	0.1313	0.1363	0.7154	0.7081
2	0.7303	0.1057	0.1116	0.8211	0.8194
3	1.0955	0.0759	0.0799	0.8970	0.8994
4	1.4606	0.0488	0.0499	0.9458	0.9498
5	1.8258	0.0282	0.0277	0.9740	0.9777
6	2.1909	0.0146	0.0133	0.9886	0.9911
7	2.5561	0.0069	0.0056	0.9955	0.9969
8	2.9212	0.0029	0.0021	0.9984	0.9990
9	3.2864	0.0011	0.0007	0.9995	0.9997
10	3.6510	0.0004	0.0002	0.9999	0.9999

TABLE 7

Higher Pearson Coefficients w.r.t. distributions defined by  $(\frac{1}{2} + \frac{1}{2})^r$ ,  $(\frac{2}{3} + \frac{1}{3})^r$ ,  $(\frac{3}{4} + \frac{1}{4})^r$  and  $(\frac{9}{10} + \frac{1}{10})^r$ .

$r =$	12				20				30			
$p =$	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{1}{4}$	$\frac{1}{10}$	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{1}{4}$	$\frac{1}{10}$	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{1}{4}$	$\frac{1}{10}$
$\beta_1$	0	0.04	0.11	0.59	0	0.03	0.06	0.36	0	0.02	0.04	0.24
$\beta_2$	2.83	2.87	2.94	3.43	2.90	2.92	2.97	3.26	2.93	2.95	2.98	3.17
$\beta_3$	0	0.39	1.05	5.88	0	0.24	0.64	3.54	0	0.16	0.43	2.36
$\beta_4$	12.61	13.20	15.20	25.89	13.54	13.90	14.99	22.16	14.02	14.26	14.89	20.40
$\beta_5$	0	3.68	10.21	57.52	0	2.37	6.47	35.64	0	1.63	4.40	23.99
$\beta_6$	73.93	90.56	119.13	310.37	85.45	96.31	114.73	253.27	91.65	99.20	111.90	204.96

If we take the Poisson distribution for  $M = 10$  as our yardstick of satisfactory fit, we may thus say that the following range of values for the Pearson coefficients of a leptokurtic binomial variate are consistent with a good fit in the same sense :

$$\begin{aligned}\beta_1 &< 0.1 ; \quad \beta_2 < 3.1 ; \quad \beta_3 < 1.1 ; \\ \beta_4 &< 17.5 ; \quad \beta_5 < 11.0 ; \quad \beta_6 < 155.\end{aligned}$$

For *symmetrical* leptokurtic distributions, the tabulated  $t$ -variate (Type VII) dealt with in 15.04 below, provides us with a standard of comparison, and we may generate symmetrical platykurtic yardstick distributions by sampling from the rectangular universe. Table 8 exhibits the close correspondence between the normal distribution and that of the total score of the 6-fold toss of a tetrahedral die with face-scores 1, 2, 3, 4. For this distribution  $\beta_2 = 2.773$ .

The sample distribution of Table 8 is referable to score totals increasing by unit steps from 6 to 24, being therefore a distribution with 19 score classes like the symmetrical binomial variate defined by successive terms of  $(\frac{1}{2} + \frac{1}{2})^{18}$ . For the latter distribution  $M = 9$  and  $\beta_2 = 2.89$ ; and we may expect a very satisfactory normal fit for a symmetrical platykurtic distribution of 20 or more score classes if  $\beta_2 \geq 2.8$ . Needless to say, fitting a continuous curve to a discrete distribution is an unduly hopeful undertaking if the number of score classes is fewer.

With due regard to the caveat last stated, the rapidity with which the mean score of the sample approaches normality is remarkable when the distribution is not very skew. Table 9 is instructive from this viewpoint. It describes the distribution of 14-fold samples from a U-shaped burette (3-class) universe, i.e. a sample distribution of 43 score classes. The normal estimate of odds against a value numerically as great as or greater than  $2.1\sigma$  is 39:1 as against 38:1 assigned by the exact distribution. For comparison Tables 10 and 11 give respectively the 10-fold sample mean-score distributions from a skew and slightly platykurtic 3-class universe and the 8-fold sample mean-score distribution from a skew and slightly leptokurtic 3-class universe.

In 14.06 we have seen that the raw score and proportionate score difference distributions of equal ( $a$ -fold) samples from an infinite 2-class universe are identical and that their moments must lie nearer to the normal than do those of the distribution of the  $2a$ -fold sample score sum or mean score except when  $m = 1$  ( $p = \frac{1}{2} = q$ ), in which case the three distributions are identical. The accompanying Table 12 exhibits how close is the normal fit for more or less skew binomial difference distributions, each being referable to the difference between the scores of samples of 10 from a 2-class universe for  $p = 0.5, 0.25, 0.20$  and  $0.10$ .



TABLE 8

*Comparison between the Normal Integral and a 6-fold toss of a Tetrahedral Die.*

$$\sigma_x = 2.739 \quad {}_4\beta_2 = 2.773$$

Score-Sum Deviation $X$	$\frac{X}{\sigma_x}$	Frequency		Cumulative Frequency	
		Exact	Normal	Exact	Normal
0	0.0000	0.1416	0.1457	0.1416	0.1449
1	0.3651	0.1333	0.1363	0.4082	0.4161
2	0.7302	0.1113	0.1116	0.6308	0.6386
3	1.0953	0.0820	0.0799	0.7948	0.7987
4	1.4604	0.0527	0.0501	0.9002	0.8996
5	1.8255	0.0293	0.0275	0.9588	0.9555
6	2.1906	0.0137	0.0137	0.9862	0.9824
7	2.5557	0.0051	0.0056	0.9964	0.9938
8	2.9208	0.0015	0.0020	0.9994	0.9981
9	3.2859	0.0003	0.0007	1.0000	0.9995

TABLE 9

*Comparison between the Normal Integral and the burette sampling distribution specified by  $(\frac{2}{5} + \frac{1}{5} + \frac{2}{5})^{14}$ .*

$$\sigma = 3.347 \quad {}_4\beta_2 = 2.875$$

Score-Sum Deviation $X$	$\frac{X}{\sigma}$	Frequency		Cumulative Frequency	
		Exact	Normal	Exact	Normal
0	0.0000	0.1173	0.1192	0.1173	0.1180
1	0.2988	0.1124	0.1140	0.3421	0.3460
2	0.5975	0.0992	0.0997	0.5405	0.5449
3	0.8963	0.0801	0.0798	0.7007	0.7043
4	1.1951	0.0589	0.0584	0.8185	0.8212
5	1.4939	0.0410	0.0391	0.9005	0.8996
6	1.7926	0.0247	0.0239	0.9499	0.9478
7	2.0914	0.0137	0.0134	0.9773	0.9750
8	2.3902	0.0068	0.0068	0.9909	0.9889
9	2.6889	0.0030	0.0032	0.9969	0.9954
10	2.9877	0.0011	0.0014	0.9991	0.9983
11	3.2865	0.0004	0.0005	0.9999	0.9994

TABLE 10

Comparison between the Normal Integral and the burette sampling distribution specified by  $(\frac{1}{6} + \frac{2}{6} + \frac{3}{6})^{10}$ .

$$\sigma = 2.357 \quad {}_t\beta_2 = 2.904$$

Score-Sum Deviation $X$	$\frac{X}{\sigma}$	Frequency		Cumulative Frequency	
		Exact	Normal	Exact	Normal
$-\frac{31}{3}$	-4.3841	0.00002	0.00001	0.00002	0.00001
$-\frac{28}{3}$	-3.9598	0.00013	0.00007	0.00015	0.00009
$-\frac{25}{3}$	-3.5356	0.00057	0.00033	0.00072	0.00044
$-\frac{22}{3}$	-3.1113	0.00203	0.00134	0.00275	0.00187
$-\frac{19}{3}$	-2.6870	0.00601	0.00458	0.00876	0.00667
$-\frac{16}{3}$	-2.2628	0.01514	0.01309	0.02390	0.02016
$-\frac{13}{3}$	-1.8385	0.03275	0.03123	0.05665	0.05195
$-\frac{10}{3}$	-1.4142	0.06106	0.06227	0.11771	0.11466
$-\frac{7}{3}$	-0.9900	0.09826	0.10368	0.21597	0.21834
$-\frac{4}{3}$	-0.5657	0.13628	0.14422	0.35225	0.36182
$-\frac{1}{3}$	-0.1414	0.16214	0.16757	0.51439	0.52820
$+\frac{2}{3}$	+0.2828	0.16413	0.16262	0.67852	0.68969
$+\frac{5}{3}$	+0.7071	0.13947	0.13442	0.81799	0.82102
$+\frac{8}{3}$	+1.1314	0.09748	0.08924	0.91547	0.91044
$+\frac{11}{3}$	+1.5556	0.05425	0.05047	0.96972	0.96144
$+\frac{14}{3}$	+1.9799	0.02279	0.02384	0.99251	0.98581
$+\frac{17}{3}$	+2.4042	0.00651	0.00941	0.99902	0.99555
$+\frac{20}{3}$	+2.8284	0.00098	0.00310	1.00000	0.99882

TABLE 11

Comparison between the Normal Integral and the burette sampling distribution specified by  $(\frac{1}{10} + \frac{7}{10} + \frac{2}{10})^8$ .

$$\sigma = 1.523 \quad {}_t\beta_2 = 3.04$$

Score-Sum Deviation $X$	$\frac{X}{\sigma}$	Frequency		Cumulative Frequency	
		Exact	Normal	Exact	Normal
$-\frac{34}{5}$	-4.4649	0.00001	0.00001	0.00001	0.00002
$-\frac{29}{5}$	-3.8083	0.00020	0.00019	0.00021	0.00025
$-\frac{24}{5}$	-3.1517	0.00185	0.00183	0.00206	0.00238
$-\frac{19}{5}$	-2.4951	0.01138	0.01165	0.01344	0.01513
$-\frac{14}{5}$	-1.8385	0.04721	0.04834	0.06065	0.06551
$-\frac{9}{5}$	-1.1819	0.13019	0.13003	0.19084	0.19667
$-\frac{4}{5}$	-0.5253	0.23196	0.22817	0.42280	0.42192
$+\frac{1}{5}$	+0.1313	0.26039	0.25969	0.68319	0.67712
$+\frac{6}{5}$	+0.7879	0.18886	0.19204	0.87205	0.86783
$+\frac{11}{5}$	+1.4445	0.09104	0.09228	0.96309	0.96186
$+\frac{16}{5}$	+2.1011	0.02954	0.02881	0.99263	0.99243
$+\frac{21}{5}$	+2.7577	0.00640	0.00585	0.99903	0.99898
$+\frac{26}{5}$	+3.4143	0.00089	0.00077	0.99992	0.99991
$+\frac{31}{5}$	+4.0709	0.00007	0.00007	0.99999	0.99999



TABLE 12

*Distributions of Raw-Score or Proportionate (Mean) Score Difference for equal 10-fold samples from more or less skew 2-class universes, the assumption being that (a) the universe is infinite, or (b) the universe is finite and sampling is subject to the replacement condition.*

$p = \frac{1}{2} = q.$					
Raw-Score Difference $D$	$\frac{D}{\sigma}$	Frequency		Cumulative Frequency	
		Exact	Normal	Exact	Normal
0	0.0000	0.1762	0.1784	0.1762	0.1770
1	0.4472	0.1602	0.1614	0.4966	0.4977
2	0.8944	0.1201	0.1194	0.7368	0.7344
3	1.3416	0.0739	0.0725	0.8846	0.8824
4	1.7889	0.0370	0.0361	0.9586	0.9558
5	2.2361	0.0148	0.0151	0.9882	0.9861
6	2.6833	0.0046	0.0049	0.9974	0.9963
7	3.1305	0.0011	0.0013	0.9996	0.9992
8	3.5777	0.0002	0.0003	1.0000	0.9999

$p = \frac{3}{4}; q = \frac{1}{4}.$					
0	0.0000	0.2056	0.2060	0.2056	0.2038
1	0.5164	0.1800	0.1801	0.5658	0.5614
2	1.0328	0.1212	0.1208	0.8080	0.8033
3	1.5492	0.0625	0.0623	0.9330	0.9293
4	2.0656	0.0247	0.0246	0.9824	0.9799
5	2.5820	0.0073	0.0074	0.9970	0.9955
6	3.0984	0.0016	0.0017	1.0000	0.9992
7	3.6148	0.0003	0.0003	1.0000	0.9999

$p = \frac{4}{5}; q = \frac{1}{5}.$					
0	0.0000	0.2238	0.2230	0.2238	0.2202
1	0.5590	0.1909	0.1907	0.6056	0.5982
2	1.1180	0.1190	0.1194	0.8436	0.8377
3	1.6770	0.0544	0.0546	0.9524	0.9496
4	2.2360	0.0184	0.0183	0.9892	0.9881
5	2.7950	0.0046	0.0045	0.9984	0.9979
6	3.3540	0.0008	0.0008	1.0000	0.9997
7	3.9130	0.0001	0.0001	1.0000	1.0000

$p = \frac{9}{10}; q = \frac{1}{10}.$					
0	0.0000	0.3126	0.2974	0.3126	0.2907
1	0.7454	0.2219	0.2252	0.7564	0.7395
2	1.4907	0.0920	0.0979	0.9404	0.9376
3	2.2361	0.0246	0.0244	0.9896	0.9909
4	2.9814	0.0045	0.0034	0.9986	0.9992
5	3.7268	0.0006	0.0004	0.9998	1.0000
6	4.4721	0.0001	0.0000	1.0000	

Tables 13 and 14 exhibit score difference distributions for samples of equal size from *burette* (3-class) universes.

TABLE 13

*Comparison between the Normal Integral and the Mean-Score Difference Distribution of 6-fold samples drawn from a Burette Universe in which  $p_a = 0.3$ ,  $h = 2$ .*

$$\sigma = 0.5196 \quad \beta_2(d) = 2.877.$$

Mean-Score Difference Deviation $d$	$\frac{d}{\sigma}$	Frequency		Cumulative Frequency	
		Exact	Normal	Exact	Normal
0	0.0000	0.13393	0.12796	0.13393	0.12746
$\frac{1}{6}$	0.3208	0.11254	0.12154	0.35901	0.36959
$\frac{2}{6}$	0.6415	0.11045	0.10415	0.57991	0.57738
$\frac{3}{6}$	0.9623	0.07570	0.08053	0.73131	0.73840
$\frac{4}{6}$	1.2830	0.06146	0.05624	0.85423	0.85107
$\frac{5}{6}$	1.6040	0.03360	0.03537	0.92143	0.92228
$\frac{6}{6}$	1.9246	0.02250	0.02008	0.96643	0.96293
$\frac{7}{6}$	2.2453	0.00939	0.01029	0.98521	0.98589
$\frac{8}{6}$	2.5661	0.00513	0.00475	0.99546	0.99358
$\frac{9}{6}$	2.8868	0.00149	0.00198	0.99846	0.99768
$\frac{10}{6}$	3.2076	0.00065	0.00075	0.99976	0.99923
$\frac{11}{6}$	3.5288	0.00010	0.00025	0.99996	0.99977
$\frac{12}{6}$	3.8492	0.00003	0.00008	1.00000	0.99994

TABLE 14

*Comparison between the Normal Integral and the Mean-Score Difference Distribution of 4-fold samples drawn from a Burette Universe in which  $p_a = 0.1$ ,  $h = 2$ .*

$$\sigma = 0.3808 \quad \beta_2(d) = 3.038$$

Mean-Score Difference Deviation $d$	$\frac{d}{\sigma}$	Frequency		Cumulative Frequency	
		Exact	Normal	Exact	Normal
0	0.0000	0.26427	0.26191	0.26427	0.25732
$\frac{1}{4}$	0.6565	0.21105	0.21111	0.68637	0.67526
$\frac{2}{4}$	1.3130	0.10933	0.11071	0.90503	0.89928
$\frac{3}{4}$	1.9695	0.03749	0.03766	0.98001	0.97844
$\frac{4}{4}$	2.6260	0.00857	0.00833	0.99715	0.99686
$\frac{5}{4}$	3.2825	0.00129	0.00120	0.99973	0.99969
$\frac{6}{4}$	3.9391	0.00012	0.00011	0.99997	0.99998
$\frac{7}{4}$	4.5956	0.00001	0.00001	0.99999	0.99999

It is very important to recognise the implications of the half interval correction when assessing the odds in favour of or against a score value equal to or greater than a standard score of unit variance definitive of a sample from a discrete universe. The variance of the distribution whose definitive binomial is  $(\frac{1}{2} + \frac{1}{2})^{16}$  is  $rpq = 4$ , so that  $\sigma = 2$ . Thus the standard score corresponding to a raw score deviation of  $+4$  ( $x = 12$ ) is 2. With due regard to the half interval, the corresponding entry in the normal table is  $(4 + \frac{1}{2}) \div 2 = 2.25$ . The vector probability that a normal score of unit variance will not exceed  $+2.25$  is about 0.9875 or odds of over 70 : 1 against.





The score-sum distribution is thus :

<i>Score</i>	.	$\frac{8}{3}$	$\frac{11}{3}$	$\frac{14}{3}$	$\frac{17}{3}$	$\frac{20}{3}$	$\frac{23}{3}$	$\frac{26}{3}$	$\frac{29}{3}$	$\frac{32}{3}$	$\frac{35}{3}$	$\frac{38}{3}$	$\frac{41}{3}$
<i>Frequency</i>	.	$\frac{1}{64}$	$\frac{1}{64}$	$\frac{5}{64}$	$\frac{5}{64}$	$\frac{10}{64}$	$\frac{10}{64}$	$\frac{10}{64}$	$\frac{10}{64}$	$\frac{5}{64}$	$\frac{5}{64}$	$\frac{1}{64}$	$\frac{1}{64}$

For the generating functions we may write

$$G_u(a) = \frac{1}{4}(e^{\frac{8h}{3}} + e^{\frac{11h}{3}} + e^{\frac{14h}{3}} + e^{\frac{17h}{3}});$$

$$G_u(b) = \frac{1}{16}(e^{\frac{h}{3}} + 4e^{\frac{5h}{3}} + 6e^{\frac{11h}{3}} + 4e^{\frac{17h}{3}} + e^{\frac{23h}{3}}).$$

By direct multiplication we obtain

$$G_u(a) \cdot G_u(b) = \frac{1}{64}(e^{\frac{8h}{3}} + e^{\frac{11h}{3}} + 5e^{\frac{14h}{3}} + 5e^{\frac{17h}{3}} + 10e^{\frac{20h}{3}} + 10e^{\frac{23h}{3}} \dots \text{etc.})$$

Each coefficient in the above corresponds to one of the frequency terms of the sum distribution, and the co-factor of  $h$  in the exponent of  $e$  is the corresponding score-sum itself on the assumption that we take 2 unit samples one from each universe. It is very important to recognise that this is not the same thing as taking a 2-fold sample from the equivalent homogeneous (*Bernoullian*) universe. Since our sample prescription is that we take equal (unit) samples from each stratum we can conceptualise a corresponding *destratified* universe on the assumptions: (a) that the two strata each contain the same number (16) of items and the common pool contains 32 in all; (b) that we replace each item before drawing another. The u.s.d. of this pooled universe is then

<i>Score</i>	.	.	$-\frac{1}{3}$	$\frac{5}{3}$	3	$\frac{11}{3}$	4	5	$\frac{17}{3}$	6	$\frac{23}{3}$
<i>Frequency</i>	.	.	$\frac{1}{32}$	$\frac{4}{32}$	$\frac{4}{32}$	$\frac{6}{32}$	$\frac{4}{32}$	$\frac{4}{32}$	$\frac{4}{32}$	$\frac{4}{32}$	$\frac{1}{32}$

For such a Bernoullian universe we may write

$$G_u = \frac{1}{32}(e^{\frac{h}{3}} + 4e^{\frac{5h}{3}} + 4e^{\frac{9h}{3}} + 6e^{\frac{11h}{3}} + 4e^{\frac{12h}{3}} + 4e^{\frac{15h}{3}} + \dots + 4e^{\frac{17h}{3}} + 4e^{\frac{18h}{3}} + e^{\frac{23h}{3}}).$$

The generating function of the 2-fold sample from the same universe is  $G_u^2$ , from which we derive by direct multiplication the following score-sum distribution as the reader may check by recourse to the grid :

<i>Score</i> ( $\times 3$ )	<i>Frequency</i> ( $\times 1024$ )	<i>Score</i> ( $\times 3$ )	<i>Frequency</i> ( $\times 1024$ )
-2	1	26	80
4	8	27	64
8	8	28	56
10	28	29	80
11	8	30	48
14	40	32	40
16	56	33	32
17	40	34	28
18	16	35	40
20	80	36	16
21	32	38	8
22	70	40	8
23	80	41	8
24	48	46	1

The foregoing example shows that : (a) the score-sum distribution w.r.t. unit samples from each of  $p$  strata of a stratified universe is very different from that of  $p$  successive samples from



To bring this assumption to life, we may suppose that each universe of the foregoing example is a lottery wheel like that of Fig. 97 assigned to one of two players and then consider the result of recording the score of a 2-fold spin of each. At each trial we thus record two scores,  $a_1, a_2$  w.r.t. lottery wheel A and  $b_1, b_2$  w.r.t. lottery wheel B. Instead of allocating to the player the score-sum, we may record the result as the paired difference scores of each of them, *viz.* :  $d_a = a_1 - a_2$  and  $d_b = b_1 - b_2$ . In 14.02 we have seen that the distribution of both sets of paired  $d$ -scores must be symmetrical about *zero mean* ; and we may write their generating function as follows :

$$G(d_b) = \frac{1}{2^5 6} \left( e^{-\frac{h}{3}} + 4e^{\frac{5h}{3}} + 6e^{\frac{11h}{3}} + 4e^{\frac{17h}{3}} + e^{\frac{23h}{3}} \right) \left( e^{\frac{h}{3}} + 4e^{-\frac{5h}{3}} + 6e^{-\frac{11h}{3}} + 4e^{-\frac{17h}{3}} + e^{-\frac{23h}{3}} \right).$$

$$G(d_a) = 2^{-4} \left( e^{-\frac{3h}{2}} + e^{-\frac{h}{2}} + e^{\frac{h}{2}} + e^{\frac{3h}{2}} \right)^2 . \quad . \quad . \quad . \quad (\text{iii})$$

$$G(d_b) = 2^{-8}(e^{-\hbar} + e^{\hbar})^8 \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (\text{iv})$$

Scores	Frequencies		
	$d_b$	$d_a$	Total
— 8	1	0	1
— 6	8	0	8
— 4	28	0	28
— 3	0	16	16
— 2	56	32	88
— 1	0	48	48
0	70	64	134
1	0	48	48
2	56	32	88
3	0	16	16
4	28	0	28
6	8	0	8
8	1	0	1
Total	256	256	512

The results of recording pair difference scores from the two lottery wheels with border-scores as specified at the beginning of this section would, of course, be the same as those of recording the single score distribution of lottery wheels with score distributions respectively identical with those of  $d_a$  and  $d_b$  above. We then have again a universe of 2 strata but this time the means of the two distributions are identical, both being zero. On the assumption that we take samples of equal size from each stratum, the equivalent homogeneous universe is as specified by equal weighting in the column marked "Total". In this composite distribution each frequency term is the sum of the corresponding terms of the stratum distributions divisible by twice the

We thus see that the distribution of the sum of  $p$  paired difference scores from different normal universes is necessarily normal, but it is equivalent to the distribution of the  $p$ -fold sample score-sum from the homogeneous pooled universe only if one assumption holds good. If the sub-universes (*strata*) from which we extract our score-pairs differ w.r.t. the value of the *mean* only, the distribution of the paired difference having zero mean will be identical from stratum to stratum, and taking  $p$  samples from any one of them amounts to the same as taking one sample from each of the  $p$  strata. If so, we can look on each  $d$ -score as a unit sample from



one and the same normal universe, and since the  $d$ -score has zero mean, an unbiased estimate of its variance ( $V_u$ ) based on  $p$  samples is

$$s_u^2 = E(d_x^2) = \frac{1}{p} \sum_{x=1}^{x=p} d_x^2 . \quad . \quad . \quad . \quad . \quad . \quad (\text{xii})$$

For the variance of the mean score of a  $p$ -fold sample from such a universe we have as our unbiased estimate

$$s_m^2 = \frac{s_u^2}{p} = \frac{1}{p^2} \sum_{x=1}^{x=p} d_x^2 . \quad . \quad . \quad . \quad . \quad . \quad (\text{xiii})$$

The mean  $d$ -score itself is

$$M = \frac{1}{p} \sum_{x=1}^{x=p} d_x.$$

Hence we derive the empirical critical ratio ( $c$ ) or standard score in the form

$$\frac{M^2}{s_m^2} \simeq c^2 \simeq \frac{\left(\sum_{x=1}^{x=p} d_x\right)^2}{\sum_{x=1}^{x=p} d_x^2} \quad . \quad . \quad . \quad . \quad . \quad (\text{xiv})$$

The justification for the derivation of (xiv) is that the  $d$ -scores constitute a homogeneous normal universe whose u.s.d. variance is definable in the usual way.

We shall now assume that the variances of the  $d$ -score distributions are not identical, in which case their  $p$ -fold score-sum distribution is still normal in accordance with (viii), the variance being

$$V = V_a + V_b + V_c + \dots \quad (xv)$$

If we denote the score-sum by  $s$ , we have  $s = pM$ , and derive the variance ( $V_m$ ) of the mean score deviation by the usual scalar transformation :

$$V = p^2 V_m. \quad . \quad . \quad . \quad . \quad . \quad . \quad (\text{xvi})$$

To evaluate  $V$  in terms of our observations, we note that we have one sample score  $d_a$  from which to estimate  $V_a$ , one sample score  $d_b$  from which to estimate  $V_b$  and so on. If we did not know that the true mean of each of the  $d$ -score distributions is zero, we could not do this; but our assumption is that we do have this knowledge. To appreciate the implications of this let us make explicit the distinction between the true mean ( $M_u$ ) and the sample mean ( $M_s$ ) of an  $r$ -fold sample of scores ( $x$ ), employing the symbol  $E_s(. . .)$  to signify the operation of extracting the mean of the complete sampling distribution. We may then write

$$E_s(\sigma_x^2) = V_u \quad \text{if} \quad \sigma_x^2 = \frac{\sum_1^r (x - M_u)^2}{r} \quad . \quad . \quad . \quad . \quad (\text{xvii})^*$$

$$E_s(s_x^2) = V_u \quad \text{if} \quad s_x^2 = \frac{\sum (x_s - M_s)^2}{r-1}. \quad (\text{xviii})$$

\* More fully, if  $E_r(\dots)$  is the operation of extracting the sample mean and  $E_s \cdot E_r \equiv E \equiv E_r \cdot E_s$ :

$$V_u = E(x - M_u)^2 = E(x^2) \text{ if } M_u = 0;$$

$$\sigma_x^2 = E_r(x - M_u)^2 = E_r(x^2) \text{ if } M_u = 0.$$

Whence if  $M_u = 0$

$$E_g(\sigma_x^2) = E_g \cdot E_r(x^2) = E(x^2) = V_u.$$

In (xviii)  $M_s = x$  if we have only one sample value of  $x$  and  $(x_s - M_s) = 0$ ; but if so  $r = 1$  and  $(r - 1) = 0$ , whence the value of  $s_x^2$  is indeterminate. If we know that  $M_u = 0$  in (xvii),  $\sigma_x^2$  is an unbiased estimate of  $V_u$ . We may thus write for our distribution of unit sample  $d$ -scores from different universes

$$E_s(d_a^2) = V_a, \quad E_s(d_b^2) = V_b \text{ etc.}, \\ \therefore E_s(d_a^2 + d_b^2 + d_c^2 \dots) = V.$$

In accordance with (xvi) our unbiased estimate of  $V_m$  is therefore

$$\frac{\sum d_x^2}{p^2}.$$

As before, we may write the square of the mean  $d$ -score as

$$M^2 = \frac{(\sum d_x)^2}{p^2}.$$

Whence we derive the empirical square critical ratio of the mean  $d$ -score, i.e. square standard mean  $d$ -score, in the same form as (xiv), viz.:

$$\frac{M^2}{V_m} = c^2 = \frac{(\sum d_x)^2}{\sum d_x^2}. \quad \dots \dots \dots \quad (\text{xix})$$

Thus the rationale of the approximate (large sample)  $c$ -test for paired differences *does not necessitate the assumption that each paired difference is a unit sample from a universe with the same variance as each other such universe.*

#### EXERCISE 14.08

1. Four players each toss one of 4 tetrahedral dice with face scores respectively as follows :

1, 2, 2, 3 ; 2, 3, 3, 4 : 3, 4, 4, 5 ; 4, 5, 5, 6.

Cite the mean and the variance of

- (a) the u.s.d. of each player's score ;
- (b) the distribution of the score-sum of each player's double toss ;
- (c) the distribution of the score difference of each player's double toss.

2. Write down the m.g.f. of each of the distributions (a)-(c) above, and specify the frequencies of possible score values.

3. Repeat Exercises 1 and 2 above for the case when the dice have face scores as follows :

1, 2, 3, 3 ; 3, 4, 5, 6 ; 3, 5, 5, 7 ; 1, 4, 4, 7.

4. Compare the results of Exercises 1-3 above, and draw your own conclusions w.r.t. what features of the u.s.d. are relevant to the character of the 2-fold toss *difference* distribution.

#### 14.09 THE USE OF PAIRED DIFFERENCES

The considerations advanced in 14.08 have a special bearing on the merits and disadvantages of pairing observations in one or other way mentioned below. In general, pairing in virtue of a similarity relevant to the end in view is always a wise procedure, if the outcome of the experiment is so clear-cut that need for statistical analysis does not arise. Otherwise, it is important to



realise that pairing presupposes sampling in a stratified universe and sampling in a stratified universe may be intractable from the statistical standpoint.

When we pair data, we conceive the universe as potentially stratified w.r.t. 2 criteria of classification, which we may specify as: (i) treatment (*columns*); (ii) within-pair resemblance (*rows*). The following possibilities then arise:

*Case 1.*

The universe is homogeneous in both dimensions ;

*Case 2.*

the universe is homogeneous in the row dimensions alone, i.e. from pair to pair for one and the same treatment ;

*Case 3.*

the universe is homogeneous in the column dimension alone, i.e. for different treatments on members of the same pair ;

*Case 4.*

the universe is homogeneous in neither dimension, in which event we have no guarantee that the composite sample will provide more than a single sub-sample for any relevant parameter.

Cases 1 and 2 are trivial in this context, since we accomplish (and lose) nothing by pairing when corresponding (within column) members of different pairs are unit samples from identical sub-universes. Within the framework of *appropriate* assumptions, Case 3 and Case 4 may each be reducible to Case 1 if we employ the method of scoring by paired differences ; but there appears to be prevalent some misconception about what the appropriate assumptions are. As regards Case 3, the relevant issue comes into focus when we examine a distinction between the following models. Each type consists of a series of different dice, each of which a player tosses twice. We thus have a pair of score values for each die and the stratified universe of which our paired scores are stratum-samples is homogeneous in one dimension in virtue of the fact that each member of a pair is a unit sample from the same stratum :

*Model I.* Four players each toss twice one of four tetrahedral dice with face-scores respectively as follows :

1, 2, 2, 3 ; 2, 3, 3, 4 ; 3, 4, 4, 5 ; 4, 5, 5, 6

The variances of the single toss distributions are the same for each die ; but the mean scores are different, *viz.* 2, 3, 4 and 5 respectively. For the 2-fold toss the variance of each player's score difference distribution is the same, being unity and the mean is zero in each case.

*Model II.* Four players each toss twice one of 4 tetrahedral dice with face-scores respectively as follows :

1, 2, 2, 3 ; 3, 4, 5, 6 ; 3, 5, 5, 7 ; 1, 4, 4, 7

The means of the single toss distributions (2, 4.5, 5, 4) are different, as are also the variances (0.5, 1.25, 2, 4.5). For the 2-fold toss, the mean of each player's difference score is the same, being zero ; but the variances are different, being respectively 1, 2.5, 4, 9.

In both series, the mean scores for the 2-fold toss are sample scores from strata with different definitive parameters ; but this is not true of the score differences as we see from the following lay-out in which  $M$  is the expected mean score of the stratum and  $\sigma_m^2$  the variance of its distribution ;  $d_m$  is the expected score difference and  $\sigma_d^2$  the variance of its distribution.

Model I				Model II			
$M$	$\sigma_m^2$	$d_m$	$\sigma_d^2$	$M$	$\sigma_m^2$	$d_m$	$\sigma_d^2$
2	0.250	0	1	2	0.250	0	1
3	0.250	0	1	4.5	0.625	0	2.5
4	0.250	0	1	5	1.0	0	4
5	0.250	0	1	4	2.25	0	9

Though we have here the variance only, it goes without saying that all the *mean* moments of the stratum  $d$ -score distributions are identical for Model 1. Thus the stratum difference score distributions are in fact identical. Since sampling from  $n$  identical strata is equivalent to sampling with replacement from any one of them or sampling from any one of them without replacement on the assumption that each stratum is indefinitely large, the Model I universe of  $d$ -scores is indeed a *homogeneous* universe.

In what circumstances we can consider Case 4 as a homogeneous universe will now be evident. Our new model will be

Model III			
Player A tosses once a tetrahedral die with face scores			
1	2	2	3
3	4	4	5
8	9	9	10
16	17	17	18
Player B tosses once a tetrahedral die with face scores			
3	4	4	5
5	6	6	7
10	11	11	12
18	19	19	20

For this model we have

$M$	$\sigma_m^2$	$d_m$	$\sigma_d^2$
3	0.25	2	1
5	0.25	2	1
10	0.25	2	1
18	0.25	2	1

The important common property of Models I and III, viz.: that the universe of  $d$ -scores is homogeneous, arises from the fact that the row-stratum distributions differ with respect to *origin* alone. This means that there is one source of *residual* variation. In terms of experimental design, such an assumption is admissible when pair to pair variation is attributable to instrumental error as in *short-term* experiments involving observations of the same subject before and after treatment. The expression short-term in this context carries with it the implication that no source of relevant individual variation obtrudes within the interval separating successive determinations. As a straightforward example of such a situation, we may consider a set of paired determinations of blood calcium level respectively carried out on different individuals immediately before injection of a fixed dose of parathyroid extract and half an hour later. If we view the issue within the traditional framework of the unique null hypothesis, our assumptions are then:

- that the scores for successive samples within the period stated differ in virtue of errors of observation only;
- this source of variation is common to all pairs of observations;
- the mean (true) value of the blood calcium level either before or after treatment is also variable from individual to individual in virtue of nature and/or nurture;
- the expected values of the blood calcium level before and after treatment are identical for one and the same individual.



A method of pairing commonly practised and commonly advocated has implications very different from the foregoing. As an example we may consider its use when the end in view is to decide whether addition of a dietetic component to the ration of a mixed population is beneficial, i.e. growth-promoting, an enquiry which typifies most situations in which the inclination to pair observations will obtrude in prophylactic or therapeutic trials. Here the observations paired are observations on different individuals chosen because they are alike w.r.t. age, sex, body-build, etc. In such circumstances we may have good reasons for believing that variance w.r.t. growth rate differs sensibly among individuals of opposite sex, of different age groups or of dissimilar physique. If so, the  $d$ -score distribution varies from row-stratum to row-stratum, and the postulates of neither Model I nor Model III hold good. In any case, it will be difficult to justify our confidence that they do so without recourse to *ad hoc* empirical data. An experiment of Cushny cited by Gossett in his original publication on the  $t$ -distribution is a somewhat unfortunate example of before-and-after-treatment pairing of observations. Here the object is to compare two optical isomers of a soporific drug by successive administration to the same individual; but the interval between the observations is inescapably protracted and presumptive major sources of variation arise less from liability to error of observation than from change of physical conditions.

## CHAPTER 15

# SAMPLING DISTRIBUTIONS

### 15.01 THE FUNDAMENTAL DISTRIBUTIONS

IN a wide range of statistical problems our concern is to explore the implications of the assumption that parameters estimated from samples are consistent with the null hypothesis that the latter come from the same universe or universes of identical structure. Such indices include sample means, sample variances, variances of sample means, the ratio of the sample mean to the sample variance, and so forth. What conclusions we can legitimately make w.r.t. their consistency presupposes knowledge of their distribution. If we can find an exact expression for the distribution of a parameter estimate, it will usually be possible to construct a table of its integral for ready reference with a view to performing a test of significance; and expressions for a wide range of such sample distributions are in fact deducible, if we assume that the normal is the u.s.d. They embrace all those which the school of R. A. Fisher invokes to test the significance of estimates of variance dealt with in Chapter 13, and of regression in Chapter 17.

Indeed, all the significance tests we shall subsequently explore rest on the assumption that the putative common universe of the null hypothesis is normal. We have already had occasion to remind ourselves that this postulate is at best a good approximation. So it is not necessary to emphasise that it is a convenient fiction. In the derivations which follow we may assume its truth regardless of its relevance to reality; and confine our attention to issues which are algebraical rather than factual. From the algebraic viewpoint, each significance test which will subsequently engage our attention is referable to one or other of a family of curves extensively investigated for the first time by Karl Pearson.

The common pattern\* which Pearson disclosed has little relevance to the end we here have in view, and the numbers he attached to the types themselves throw no light on their place in modern sampling theory. What is more important from our viewpoint is: (a) the relation in which they stand to the normal curve, as indicated symbolically in Fig. 109; (b) the fact that it is possible to define their properties uniquely in terms of the first 4 moments. In so far as our concern is with their role in sampling theory, we may summarise the familial relationship of the relevant types as follows:

- (i) With appropriate choice of constants, Type III describes the distribution of the sum of  $n$  independent *square* normal scores of zero mean, and is therefore of special relevance to the specification of the distribution of sampling variance, as set forth in 16.01–16.03;
- (ii) Type VI describes the distribution of the *ratio* of two independent Type III variates, and is therefore of special relevance to the construction of a test (the  $F$ -test) for consistency between independent estimates of variance;
- (iii) Type VII (on which the  $t$ -test mentioned in Chapter 7 of Vol. I relies) describes the distribution of the square root of a particular Type VI variate, and therefore stands in the same relation to the latter as does the normal distribution to the simplest case (Chi-Square for 1 d.f.) of Type III;
- (iv) Type I describes the ratio of a Type III variate A to the sum of two independent Type III variates A and B. It is a good descriptive curve for the distribution of the raw score of large samples drawn from a 2-class universe without replacement;

\* Pearson developed the equation of a distribution embracing all his Types as special cases from consideration of sampling without replacement from a finite 2-class universe, and placed Type I at the head of the list for that reason.



- (v) Type II is merely the symmetrical form Type I assumes as a particular case and is a good descriptive curve for the sample distribution of the product moment coefficient  $r$  from a bivariate normal universe, when its true value is zero.

From the foregoing remarks it is evident that the kingpin of the system is Type III of which the Chi-Square distribution is a special case. We have already foreshadowed its relation to the normal distribution in 14.04. The remaining types are derivable, if we can express the distribution of one score which is a function of another in terms of the distribution of the latter. This will be our preliminary task in 15.03. First, however, we may usefully recall and elaborate previous remarks on what we mean by a variate.

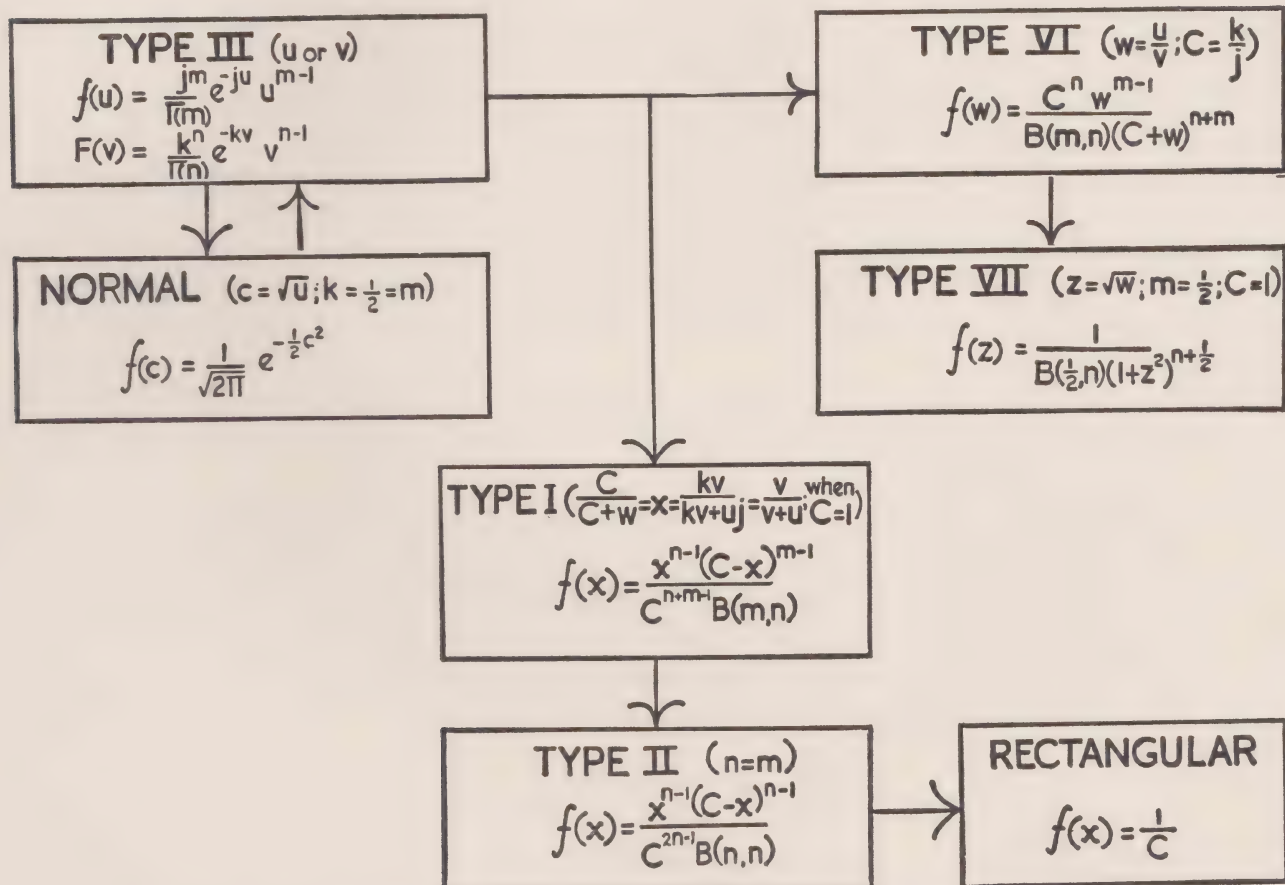


FIG. 109. The Family Tree of the Pearson System. (For the rectangular case,  $n=1$ .)

When we speak of a sample distribution, we presuppose the existence of a set of scores (*univariate* distribution) or of more than one such set (*multivariate* distribution) which constitute what we customarily speak of as the independent variable or variables. In this and the next chapter our concern will be only with univariate distributions, in which case what we refer to generically as the distribution itself is an expression connecting a particular value ( $x$ ) of the score (*variate*) with a particular value of a variable  $y = F(x)$  (*probability density*) denoting the expected frequency of score values within a specifiable range including  $x$  itself. Thus  $y$  is what we ordinarily call a dependent variable. One reason for using the terms *variate* and *probability density* is that the words dependent and independent do not have the same meaning in statistics as in co-ordinate geometry or algebra.

When we are talking about the real world, it is highly relevant to be clear about whether our scores increase by finite steps (discrete distribution) or otherwise. If they do so, a continuous





- $$\sum_{u=-a}^{u=+b} F(u) \Delta u = 1 = \sum_{u=-a}^{u=+b} \Delta A_u \quad . \quad . \quad . \quad . \quad (\text{iii})$$

(vii) To define the frequency of a score value explicitly the appropriate expression must therefore include as a factor the increment by which it increases. If we postulate that it increases *continuously* from  $a$  to  $b$ , (iii) will thus take the form

$$dA = f(u)du \quad \text{and} \quad \int_a^b f(u)du = 1 \quad . \quad . \quad . \quad . \quad (\text{iv})$$

We may sum up what has gone before as follows :

- Since our first approach to a continuous sampling distribution is the derivation of the normal curve, it is of special importance to be clear about what we mean by a normal variate. We can derive the normal curve as the limiting contour of the histogram of  $(q + p)^r$  when  $r$  is indefinitely large. If so, we can express the frequency  $F(X)\Delta X$  of the deviation  $X$  of the raw-score ( $x$ ) from its mean  $M_x = rp$  by the approximate relation :

$$F(X)\Delta X \simeq (2\pi V_x)^{-\frac{1}{2}} \cdot \exp \left[ -\frac{X^2}{2V_x} \right] \Delta X \quad . \quad . \quad . \quad . \quad (\text{v})$$

$$f_x \simeq \frac{1}{(2\pi V_x)^{\frac{1}{2}}} \exp\left(-\frac{x^2}{2V_x}\right) \simeq Y_x \quad . \quad . \quad . \quad . \quad (\text{vi})$$

This is so because  $X$  increases by unit steps from  $-M_x = -rp$  to  $(r - M_x) = rq$  as  $x$  itself increases by unit steps from 0 to  $r$ . The same successive frequencies which define  $x = 0, 1, 2 \dots r$  also specify proportionate score values :

$$0, \frac{1}{r}, \frac{2}{r} \dots 1.$$

The deviation ( $U = u - p$ ) of the proportionate score ( $u$ ) from its mean value ( $p$ ) thus increases by steps  $\Delta U = r^{-1}$  from  $-p$  to  $q$ . If we substitute  $\Delta U = r^{-1}$ ,  $rU = X$  and  $r^2 V_u = rpq = V_x$  on the right of (vi) above, it becomes

$$f_x \simeq (2\pi r^2 V_u)^{-\frac{1}{2}} \exp \left( -\frac{U^2}{2V_u} \right).$$

The frequency ( $f_x$ ) of the raw-score deviation ( $X = rU$ ) is identical with that ( $f_U$ ) of the corresponding proportionate score deviation ( $U$ ) so that

$$f_U \simeq (2\pi r^2 V_u)^{-\frac{1}{2}} \exp \left( -\frac{U^2}{2V_u} \right) \quad \dots \dots \dots \quad \text{(vii)}$$

The form of (vii) is not identical with that of (vi), but we may disclose the sense in which we can properly say that the normal is a good descriptive function for both variates ( $X$  and  $U$ ) when we express  $f_U$  in terms of the ordinate ( $Y_U$ ) and base ( $\Delta U$ ) of the corresponding column of the histogram for the  $U$ -score distribution, i.e.

$$Y_U \cdot \Delta u = f_U = r^{-1} \cdot Y_U.$$

Whence, from (vii)

$$Y_U = (2\pi V_u)^{-\frac{1}{2}} \exp \left( -\frac{U^2}{2V_u} \right) \quad \dots \dots \dots \quad \text{(viii)}$$

The last equation is formally identical with (v), i.e.  $F(U) = F(X)$ . Thus the frequency equations of 2 nearly normal variates need not be formally identical. We speak of a variate  $A$  of zero mean as normal if the ordinate (p.d.) equation is normal, i.e.

$$F(A) = (2\pi V_A)^{-\frac{1}{2}} \exp \left( -\frac{A^2}{2V_A} \right).$$

Similarly, we label a variate as a function of any type by the name of the function which defines the p.d. Thus we speak of  $x$  as a Gamma variate when we mean that the ordinate  $y$  of the distribution is expressible as the integrand of a Gamma function (see p. 246). This is at first confusing, since it is an inversion of the more common practice of attaching a verbal description to the *dependent* variable, e.g.  $y$  is a parabolic function of  $x$  if  $y = Ax^2 + B$ .

The preceding comparison between the build-up of the histogram for the raw-score and that of the proportionate score of the  $r$ -fold sample distribution from a 2-class universe which is both infinite and discrete offers a clue to the problem of defining the p.d.  $f(U)$  of a function  $U = \phi(X)$  of a continuous variate  $X$  when we know  $F(X)$ , the p.d. of  $X$  itself. If there is only one value of  $U$  for each value of  $X$  and *vice versa*, the rule for a discrete variate is implicit in the build up of the frequency histogram, *viz.* :

$$F(X)\Delta X = f(U)\Delta U \quad \dots \dots \dots \quad \text{(ix)}$$

$$\therefore f(U) = F(X) \frac{\Delta X}{\Delta U}.$$



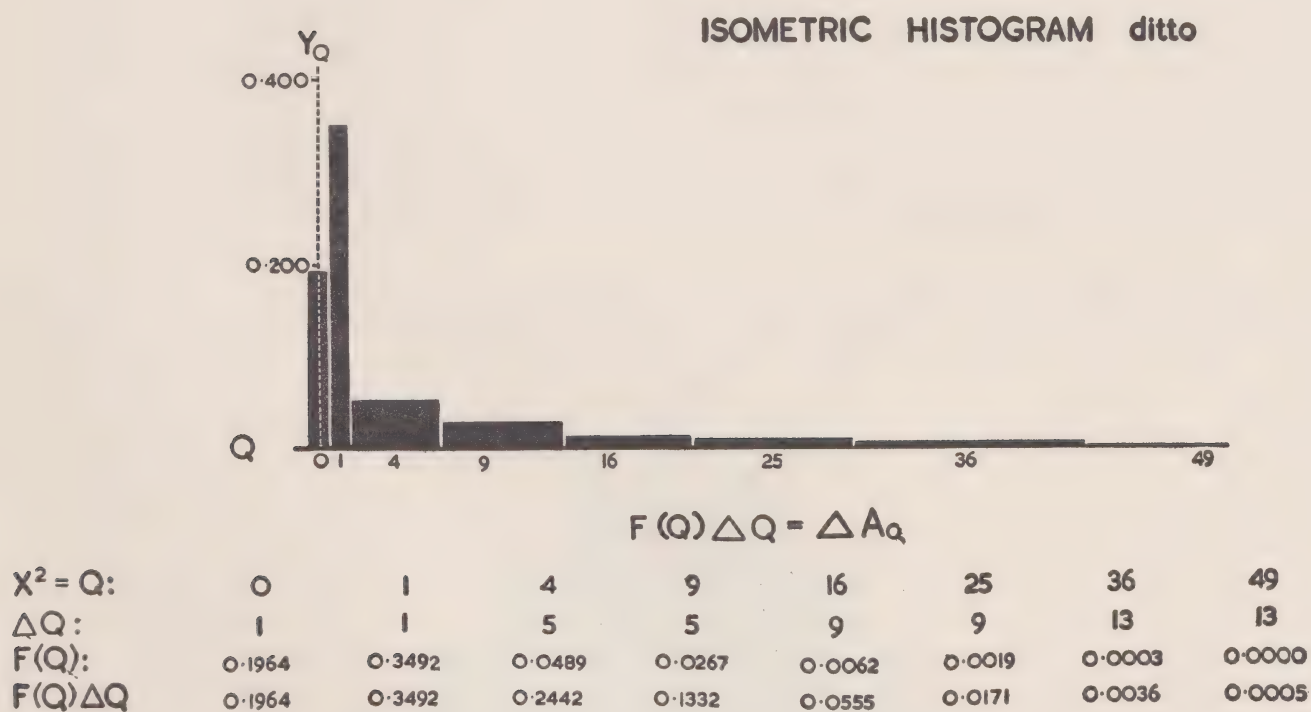
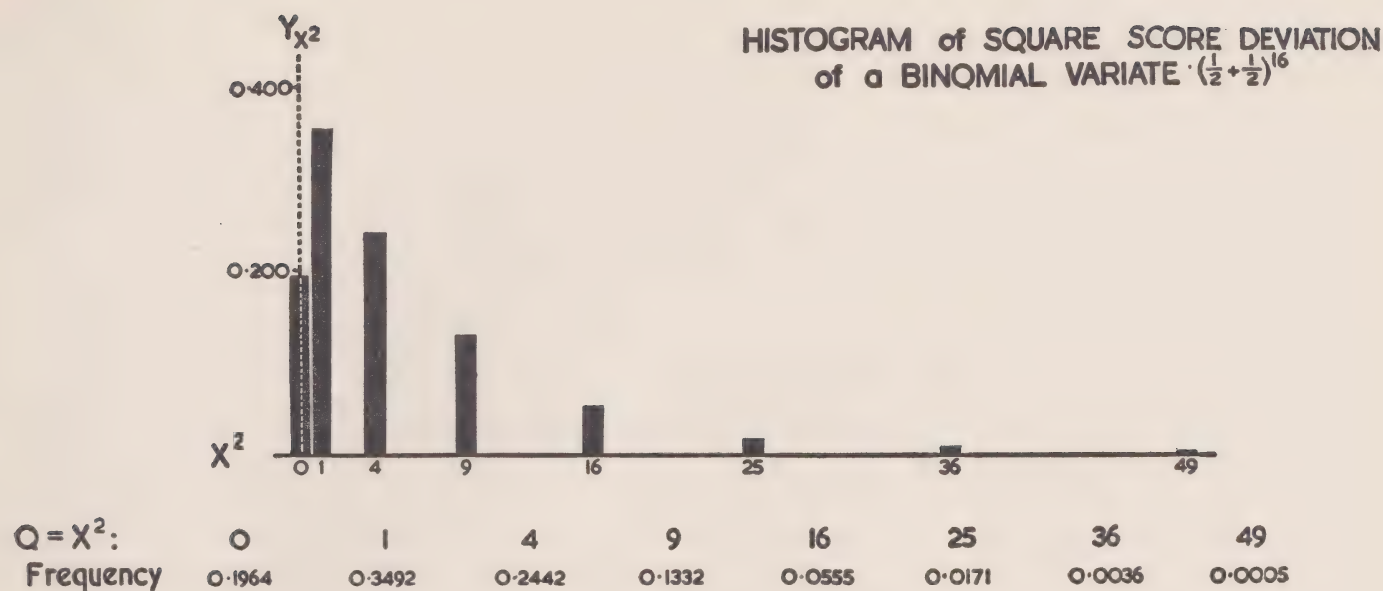


FIG. 110. Histogram of the Raw-Score deviation of the 16-fold sample from a symmetrical ( $p = \frac{1}{2} = q$ ) 2-class universe.

In the limit, this is equivalent to the following simple rule when the condition last stated\* holds good, *viz.* :

$$F(X) \frac{dX}{dU} = f(U) \quad . \quad . \quad . \quad . \quad . \quad . \quad (x)$$

\* The reader will probably recognise (x) as the principle implicit in the familiar trick of *integration by substitution* ; but few elementary textbooks on the infinitesimal calculus emphasise or illustrate, if indeed they mention, that the substitution is not always permissible in precisely this form.

In the foregoing example,

$$U = \phi(X) = \frac{X}{r}.$$

Thus one value of  $X$  corresponds to one value of  $U$  and *vice versa* throughout the whole score range negative and positive. It will be easy to see that this condition is highly relevant if we set ourselves the task of visualising the distribution of the *square* raw-score deviation  $Q = (\pm X)^2$  of the  $r$ -fold sample from the infinite 2-class universe.

If  $p = \frac{1}{2} = q$  and  $r = 16$ , as in Fig. 110,  $X$  has zero mean and 17 discrete values  $0, \pm 1, \pm 2, \pm 3, \pm 4, \pm 5, \pm 6, \pm 7, \pm 8$  with frequencies defined by appropriate terms of  $(\frac{1}{2} + \frac{1}{2})^{16}$ , viz.,  $16_{(8)} \cdot 2^{-16}, 16_{(9)} \cdot 2^{-16}, 16_{(10)} \cdot 2^{-16} \dots 16_{(16)} \cdot 2^{-16}$ .  $Q$  has 9 discrete values all positive  $0, 1, 4, 9, 16, 25, 36, 49, 64$  with frequencies defined by  $16_{(8)} \cdot 2^{-16}, 2 \cdot 16_{(9)} \cdot 2^{-16}, 2 \cdot 16_{(10)} \cdot 2^{-16} \dots 2 \cdot 16_{(16)} \cdot 2^{-16}$ . Thus the frequency of the score  $X = -3$  is  $16_{(5)} \cdot 2^{-16} = 0.0666 = 16_{(11)} \cdot 2^{-16}$  which is also the frequency of the score  $X = +3$ ; and the frequency of the score  $Q = 9 = (\pm 3)^2$  is therefore  $2(0.0666) = 0.1332$ . We can represent the frequency of each  $Q$  score other than  $Q = 0$  as a rectangular column on unit base, if we make the height equal to  $2F(X)$ , since  $Q = (\pm X)^2$ . When  $Q = 0$ , the appropriate height will be  $F(0)$ . Such a procedure will leave increasing gaps between successive columns from  $Q = 1$  onwards as in the upper half of Fig. 111. If we were to draw a smooth curve closely following the contour of the frequency polygon formed by joining the mid-points at the head of each column, the area of any segment except in the interval  $Q = 0$  to  $Q = 1$  would therefore include that of empty spaces, and would greatly exceed the sum of the frequencies of score values in the range cut off by it.

It is possible to make a histogram of the  $Q$ -score distribution *uniformly dense* (i.e. without gaps), if we abandon the luxury of making the columns of equal width ( $\Delta Q$ ) without relinquishing the two fundamental conventions of the histogram of the  $X$ -score distribution, viz.: (a) the

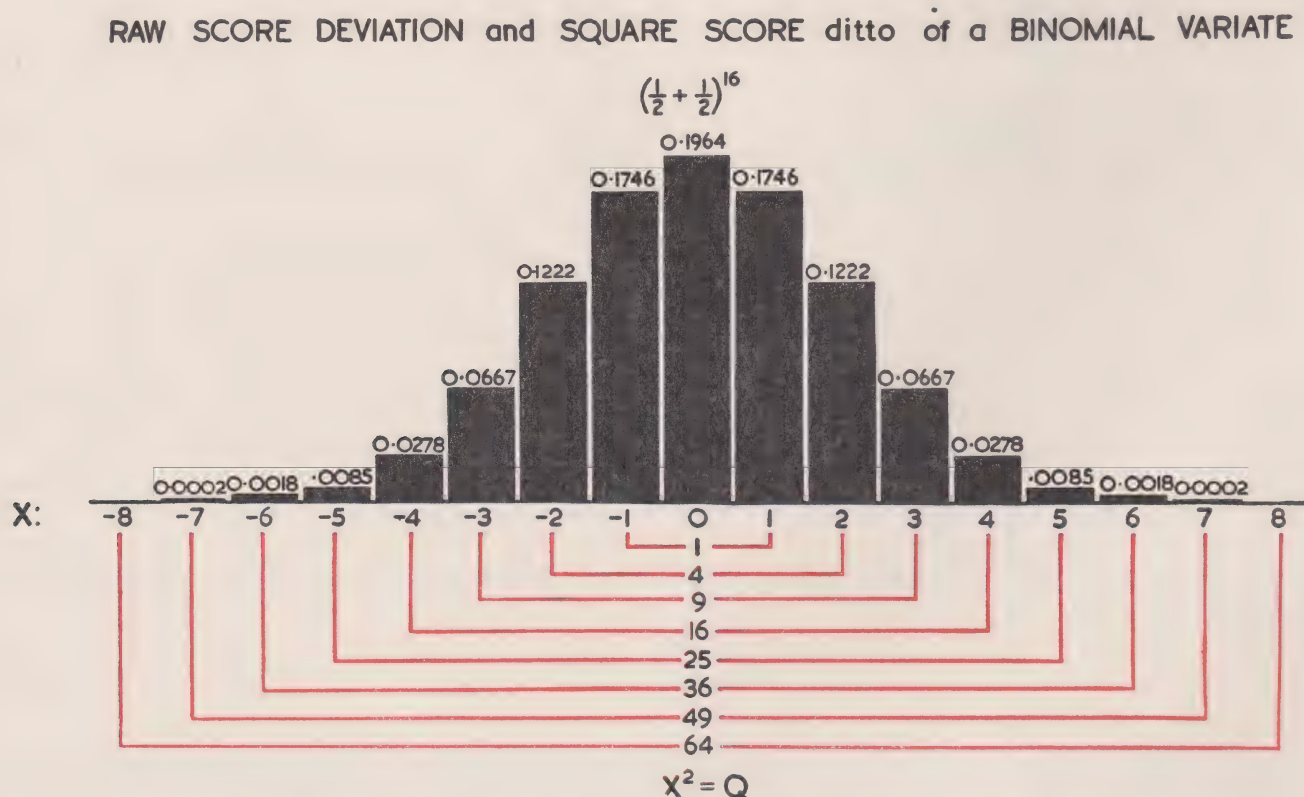


FIG. 111. Visualisation of the Square Score whose distribution is in Fig. 110.



mid-point of the base of each column specifies an actual value of  $Q$ ; (b) the area of the column specifies the frequency of the corresponding  $Q$ -score.

The fact that the first two  $Q$  scores ( $Q = 0$  and  $1$ ) are consecutive dictates how we space the boundaries of the columns in conformity with (a). The boundaries of the columns representing these two values are  $-\frac{1}{2}$  to  $+\frac{1}{2}$  and  $+\frac{1}{2}$  to  $+\frac{3}{2}$ . Thus the half width of the next column ( $Q = 4$ ) will be  $(4 - 1\frac{1}{2}) = 2\frac{1}{2}$  so that  $\Delta Q_4 = 5$ . The base of this column ends at  $(4 + 2\frac{1}{2}) = 6\frac{1}{2}$  and the half width of the next column ( $Q = 9$ ) will be  $(9 - 6\frac{1}{2}) = 2\frac{1}{2}$ , so that  $\Delta Q_9 = 5$ . By the same token, the half width of the next column ( $Q = 16$ ) will be  $(16 - 11\frac{1}{2}) = 4\frac{1}{2}$ , so that  $\Delta Q_{16} = 9$  and so on.

Having spaced our columns so that the boundaries of each are equidistant from the point which marks the corresponding  $Q$ -score on the base line, we can fulfil our second condition by defining the height  $f(Q)$  of each column in terms of  $F(X)$  accordingly. Thus  $f(0) = F(0)$ ,  $f(1) = 2F(1)$ ,  $f(4)\Delta Q_4 = 2F(2)$ ,  $f(9)\Delta Q_9 = 2F(3)$ , etc. For instance,

$$f(16)\Delta Q_{16} = 2F(4) = 0.0556,$$

$$\therefore f(16) = \frac{1}{9}(0.0556) = 0.0062.$$

The values of  $\Delta Q$  obtained in this way assume a more suggestive aspect if we place them side by side with the corresponding values of  $X$ :

$\pm X = 0$	1	2	3	4	5	6	7	8
$\Delta Q = 1$	1	5	5	9	9	13	13	17

If we now put the *mean* values of  $\Delta Q$  under the mean value of *successive pairs* of  $X$ -scores we get:

$\pm X_m = \frac{1}{2}$	$1\frac{1}{2}$	$2\frac{1}{2}$	$3\frac{1}{2}$	$4\frac{1}{2}$	$5\frac{1}{2}$	$6\frac{1}{2}$	$7\frac{1}{2}$
$\Delta Q_m = 1$	3	5	7	9	11	13	15.

We now note that  $\Delta Q_m = 2X_m$ , just as

$$\frac{dQ}{dX} = 2X.$$

Such is the build up of the isometric histogram in the lower half of Fig. 111. Let us now ask what it suggests. By removing the restrictions that  $\Delta Q$  must have a fixed value we have been able to eliminate empty spaces between columns, thus leaving open the possibility of expressing the sum of successive  $Q$ -score frequencies in terms of the area of a segment of a suitable fitting curve. Except when  $Q = 0$ , the following relation then expresses the height of the columns of the parent and the derived histogram:

$$2F(X)\Delta X = f(Q)\Delta Q,$$

$$\therefore f(Q) = 2F(X)\frac{\Delta X}{\Delta Q}.$$

The outcome suggests the following identity in the limit

$$f(Q) = 2F(X)\frac{dX}{dQ} \quad \dots \quad (xi)$$

This expression is not identical with (ix), because each value of  $Q$  corresponds to two values of  $X$  (other than  $X = 0$ ), and the frequency of any value of  $Q$  (other than zero) is therefore the sum of 2  $X$ -score frequencies, in this case equal because  $F(X)$  is a symmetrical function of  $X$  itself, being a term of the expansion of  $(\frac{1}{2} + \frac{1}{2})^{16}$ . If this is not so, as when the definitive binomial of the distribution is  $(\frac{3}{4} + \frac{1}{4})^r$ , we should write

$$f(Q)\Delta Q = F(-X)\Delta X + F(X)\Delta X \quad \dots \quad (xii)$$





*Example (I).*—The standard score is  $c = (X \div \sigma)$  and

$$F(X) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{X^2}{2\sigma^2}\right).$$

In this case  $k^{-1} = \sigma$  and (i) becomes

$$f(c) = \sigma F(\sigma c) = (2\pi)^{-\frac{1}{2}} e^{-\frac{1}{2}c^2}.$$

*Case II.* The variate  $z$  is a single-valued monotonic function of  $x$  decreasing as  $x$  increases. This case also falls into line with (x) of 15.02, since  $z$  has one value for each value of  $x$  and *vice versa*. To make clear what we are doing when we use the tabulated functions to specify a segment of area, we shall assume that  $z = a$  when  $x = c$  and  $z = b$  when  $x = d$ :

*Example (2).*—Let  $x$  be a score whose p.d. is that of a Beta variate of unlimited range, i.e. Pearson's Type VI, defined by

$$F(x) = \frac{1}{B(j, k)} \cdot \frac{x^{j-1}}{(1+x)^{k+j}} \quad . \quad . \quad . \quad . \quad . \quad . \quad (\text{ii})$$

We wish to determine  $f(z)$  the p.d. of a score defined by

$$z = \frac{1}{1+x} \quad . \quad . \quad . \quad . \quad . \quad . \quad (iii)$$

As before,  $F(x)dx$  is numerically equal to  $f(z)dz$ ; but the fact that  $x$  increases while  $z$  diminishes means that  $c > d$  if  $a < b$ , since  $b = (1 + d)^{-1}$  and  $a = (1 + c)^{-1}$ . This means that for *all* values of  $d$  or  $c$  and corresponding values of  $b$  and  $a$ ,

$$\int_a^b f(z)dz = - \int_c^d F(x)dx.$$

To interpret the sign of the integral correctly we have therefore to put

$$-F(x)\frac{dx}{dz} = f(z).$$

In accordance with (iii)

$$(1+x)^{k+j} = z^{-k-j} \quad \text{and} \quad x^{j-1} = \frac{(1-z)^{j-1}}{z^{j-1}};$$

$$\frac{dx}{dz} = \frac{d}{dz} \left( \frac{1}{z} - 1 \right) = -z^{-2}.$$

Whence we have

$$F(x) \cdot \frac{dx}{dz} = \frac{-z^{k-1}(1-z)^{j-1}}{B(k, j)}.$$

We may therefore write

$$f(z) = \frac{z^{k-1}(1-z)^{j-1}}{B(j,k)} \quad . \quad . \quad . \quad . \quad . \quad . \quad (\text{iv})$$

This is an important result, since  $f(z)$  is a Beta variate of unit range, being a case of Pearson's Type I.

*Case III.* The parent score  $x$  has a symmetrical distribution with two values corresponding to one value of  $u$ , as when  $u = (\pm x)^2$ . We shall assume that the range of  $x$  is from  $-\infty$  to  $+\infty$ , so that the range of  $u$  is from 0 to  $\infty$ . Thus  $x$  will lie in the range from  $-d$  to  $-c$  and from  $+c$  to  $+d$  when  $u$  lies in the range from  $a = (\pm c)^2$  to  $b = (\pm d)^2$ . This is the situation to which (xiii) of 15.02 applies. For any values of  $c$  and  $d$  with corresponding values of  $a$  and  $b$ ,

$$\int_{-d}^{-c} F(x)dx + \int_c^d F(x)dx = \int_a^b f(u)du.$$





*Case IV.* We now suppose that  $u$  has two values for each value of a parent variate  $x$  confining ourselves to the special case  $u = \pm \sqrt{x}$ . For real values of  $u$  that of  $x$  must be positive and we assume that its range is from 0 to  $\infty$ , so that the range of  $u$  is from  $-\infty$  to  $+\infty$ , whence this is the case covered by (xiv) of 15.02, i.e.

$$\int_c^d F(x)dx = \int_{-b}^{-a} f(u)du + \int_a^b f(u)du.$$

If we know that  $f(u)$  is a symmetrical function of  $u$ , (xv) of 15.02 holds good, i.e.

$$\int_c^d \frac{1}{2}F(x)dx = \int_a^b f(u)du.$$

When  $u = \pm \sqrt{x}$  so that  $x = u^2$

$$\frac{1}{2}F(x) \frac{dx}{du} = u \cdot F(x),$$

$$\therefore f(u) = u \cdot F(x) = x^{\frac{1}{2}} F(x) \quad \dots \dots \dots (ix)$$

*Example (4).*—We may reverse the procedure of the last example to obtain from (viii) the normal distribution of unit variance as that of the square root of the Chi-Square variate ( $C = c^2$ ) for 1 degree of freedom. In accordance with (ix)

$$f(c) = c \cdot F(C).$$

In this expression

$$F(C) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}C} \cdot C^{-\frac{1}{2}} = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}c^2} \cdot c^{-1},$$

$$\therefore c \cdot F(C) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}c^2} = f(c).$$

*Example (5).*—An analogous transformation involves the relation between Pearson's Types VI and VII. We define the former in the range  $x = 0$  to  $x = \infty$  as the p.d. of a score  $x$  such that for positive values of  $j$  and  $k$ ,

$$F(x) = \frac{1}{B(j, k)} \frac{x^{j-1}}{(1+x)^{k+j}} \quad \dots \dots \dots (x)$$

A case of special interest arises when  $j = \frac{1}{2}$ , so that

$$F(x) = \frac{1}{B(\frac{1}{2}, k)} \cdot \frac{x^{-\frac{1}{2}}}{(1+x)^{k+\frac{1}{2}}} \quad \dots \dots \dots (xi)$$

For reasons which we shall mention later this describes the distribution of a statistic whose square root ( $u = \sqrt{x}$ ) has a symmetrical distribution. Whence, in virtue of (ix),

$$u \cdot F(x) = \frac{1}{B(\frac{1}{2}, k)} \cdot \frac{ux^{-\frac{1}{2}}}{(1+x)^{k+\frac{1}{2}}}.$$

By substituting  $x = u^2$ , we thus obtain

$$f(u) = \frac{1}{B(\frac{1}{2}, k)(1+u^2)^{k+\frac{1}{2}}} \quad \dots \dots \dots (xii)$$

This is Pearson's Type VII. By hypothesis the range of  $u$  in (xii) is from  $-\infty$  to  $+\infty$  and  $f(u)$  is a symmetrical function being equal for numerically identical positive and negative values of  $u$ . Hence the mean value of  $u$  is zero. By differentiating we obtain

$$D_u \cdot f(u) = \frac{(k + \frac{1}{2})}{B(\frac{1}{2}, k)} \cdot \frac{-2u}{(1+u^2)^{k+\frac{3}{2}}}.$$











We now recall (xxx) in 11.02, *viz.* :

$$\begin{aligned} m_2 &= \mu_2 - \mu_1^2; \\ m_3 &= \mu_3 - 3\mu_2 \cdot \mu_1 + 2\mu_1^3; \\ m_4 &= \mu_4 - 4\mu_3\mu_1 + 6\mu_2 \cdot \mu_1^2 - 3\mu_1^4. \end{aligned}$$

Whence we derive

$$\therefore \beta_1 = \frac{4}{n}; \beta_2 = 3 + \frac{6}{n} = 3(1 + \frac{1}{2}\beta_1) \quad \dots \quad \text{(xii)}$$

From (xii) we see that  $\beta_1$  suffices to define  $\beta_2$ , and that  $\beta_2 > 3$ , and the distribution is both skew and leptokurtic. When  $n$  is very large, the first two Pearson coefficients do not sensibly differ from those of the normal curve. By differentiating to obtain the maximum value of  $f(x)$ , we can obtain the distance of the mode from the origin :

$$D_x f(x) = (n-1)e^{-kx} \cdot x^{n-2} - ke^{-kx} \cdot x^{n-1}.$$

Whence  $D_x f(x) = 0$  when

$$x = \frac{n-1}{k} \quad . \quad . \quad . \quad . \quad . \quad . \quad (\text{xiii})$$

The distance between the mode and mean is

$$\frac{n-1}{k} - \mu_1 = \frac{n-1}{k} - \frac{n}{k} = -k^{-1} \quad . \quad . \quad . \quad . \quad (\text{xiv})$$

For the distribution we have elsewhere (p. 644) called Chi-Square for 1 d.f.,  $k = \frac{1}{2} = n$  and  $x$  in (xiii) is negative, i.e. there is no mode in the positive range, the curve being *monotonic*. The Type III distribution is indeed *unimodal* only if  $n > 1$ . From (xiv) we then see that the mode is to the left of the mean, the distribution being therefore most steep on the side nearest the origin. More generally, we speak of the Type III distribution defined by (vi) above as the Chi-Square distribution for  $2n$  degrees of freedom when  $k = \frac{1}{2}$ . Thus Chi-Square for 2 d.f. is also monotonic, since  $n = 1$  and  $x = 0$  in (xiii); but Chi-Square for 3 d.f. has a maximum value in the positive range, since  $n = \frac{3}{2}$  and  $(n - 1) = \frac{1}{2}$ .

Evidently, the relation  $\beta_2 = 3(1 + \frac{1}{2}\beta_1)$  restricts the suitability of the Type III distribution as a good fitting curve for a distribution which is both skew and leptokurtic. Its special interest for our purpose arises from the circumstance that it includes the Chi-Square distributions as a special case when  $k = \frac{1}{2}$ ; and the importance of the latter resides in the fact that they describe the distribution of the square deviation (Chi-Square for 1 d.f.) of a normal variate and (as we shall later see) for the sum of  $f$  independent square normal score deviations of unit variance (Chi-Square for  $f$  degrees of freedom). Hence it is of fundamental relevance to statistical problems involving the distribution of variance estimates.

Since  $f = 2n$  by definition, we write Type III in Chi-Square form as

$$f(x) = \frac{2^{-\frac{1}{2}f} \cdot e^{-\frac{1}{2}x} \cdot x^{\frac{1}{2}(f-2)}}{\Gamma(\frac{1}{2}f)}.$$

From (vii) we obtain the zero moments of the distribution as

$$\mu_1 = f; \mu_2 = f(f+2); \mu_3 = f(f+2)(f+4); \mu_4 = f(f+2)(f+4)(f+6), \text{ etc.}$$

*Types I and II.* The complete Beta function is expressible\* alternatively as an integral of limited or unlimited range, *viz.* :

$$\int_0^1 x^{j-1}(1-x)^{k-1}dx = B(j, k) = \int_0^\infty \frac{x^{j-1}}{(1+x)^{k+j}}dx \quad . \quad . \quad . \quad (xv)$$

The Beta function  $B(j, k)$  is expressible in terms of the Gamma function, *viz.* :

$$B(j, k) = \frac{\Gamma(j) \Gamma(k)}{\Gamma(j+k)}.$$

We may adapt as follows the left-hand expression of (xv) to define a variate of restricted range from 0 to  $a$

$$\int_0^a x^{j-1}(a-x)^{k-1}dx = a^{j+k-1}B(j, k) \quad . \quad . \quad . \quad (xvi)$$

Accordingly, we may define a Beta variate of restricted range by the equation

$$f(x) = \frac{x^{j-1}(a-x)^{k-1}}{a^{j+k-1} \cdot B(j, k)} \quad . \quad . \quad . \quad . \quad (xvii)$$

Equation (xvii) defines Pearson's Type I of which Type II is a special case when  $j = k$ , so that

$$f(x) = \frac{x^{j-1}(a-x)^{j-1}}{a^{2j-1}B(j, j)} \quad . \quad . \quad . \quad . \quad (xviii)$$

The general expression for the zero moments is

$$\begin{aligned} \mu_r &= \int_0^a \frac{x^{j+r-1} \cdot (a-x)^{k-1} \cdot dx}{a^{j+k-1} \cdot B(j, k)} ; \\ \mu_r &= \frac{a^{j+k+r-1} \Gamma(j+r) \Gamma(j+k)}{a^{j+k-1} \Gamma(j) \Gamma(j+k+r)} = \frac{a^r (j+r-1)^{(r)}}{(j+k+r-1)^{(r)}} \quad . \quad . \quad . \quad (xix) \end{aligned}$$

Whence the mean of the distribution is

$$\mu_1 = \frac{aj}{j+k}.$$

On differentiating  $f(x)$  in (xvii) we get

$$D_x \cdot f(x) = \frac{(j-1)x^{j-2}(a-x)^{k-1} - (k-1)x^{j-1}(a-x)^{k-2}}{a^{j+k-1} \cdot B(j, k)}.$$

Hence  $D_x \cdot f(x) = 0$  when

$$x = \frac{a(j-1)}{j+k-2}.$$

This defines a turning point (*maximum*) within the prescribed positive range if  $j$  and  $k$ , being greater than unity, are (as we here assume) of the same sign, *viz.* positive. The score value ( $x$ ) so defined is then that at the mode and  $(x - \mu_1)$  is the distance of the mode from the mean, so that

$$x - \mu_1 = \frac{a(j-1)}{j+k-2} - \frac{aj}{j+k} = \frac{a(j-k)}{(j+k)(j+k-2)}.$$

\*  $j$  and  $k$  being positive.





From (xx) it is evident that Type II describes a *platykurtic* distribution ( $\beta_2 < 3$ ) for all positive values of  $j$ ; and that the values of the first two Pearson coefficients are identical with those of the normal when  $j$  is indefinitely large. We shall explore its relation to Type III on p. 658 after we have examined the properties of Type VI.

From (xix) we obtain in the usual way the following values for the mean moments of the more general (Type I) distribution :

$$m_2 = \frac{a^2kj}{(k+j)(k+j+1)^{(2)}}; \quad m_3 = \frac{2a^3kj(k-j)}{(k+j)^2(k+j+2)^{(3)}};$$

$$m_4 = \frac{3a^4jk[jk(j+k+2) + 2(k-j)^2]}{(k+j+3)^{(4)}(k+j)^3}.$$

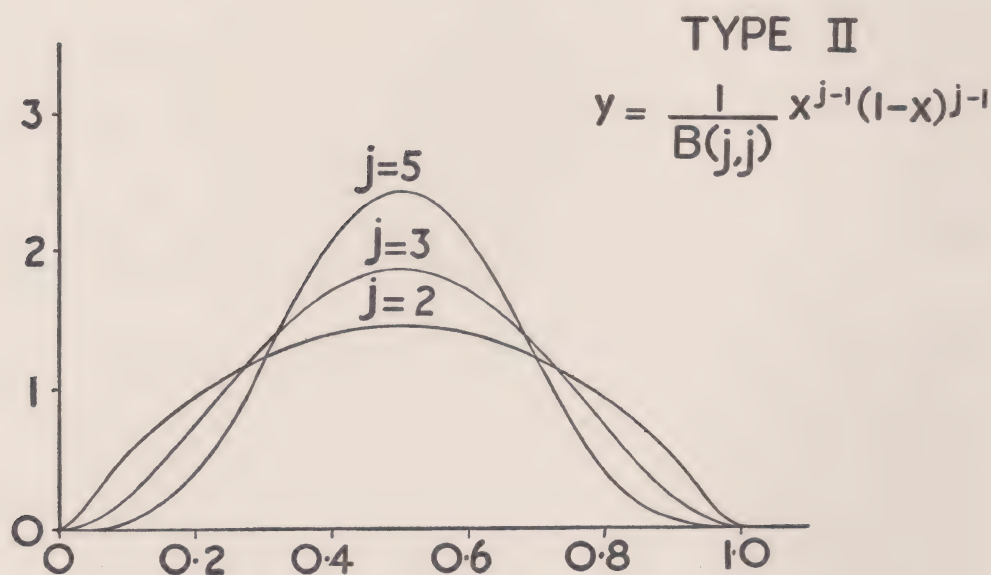


FIG. 114. Pearson's Type II.

Whence we have

$$\beta_1 = \frac{4(k-j)^2(k+j+1)}{kj(k+j+2)^2};$$

$$\beta_2 = 3 - \frac{6[kj(k+j+2) - (k-j)^2(k+j+1)]}{kj(k+j+3)^{(2)}}.$$

For positive values of  $k$  and  $j$  greater than unity, we have seen that Type I is a unimodal distribution. It may then be platykurtic, but will be leptokurtic if very skew. The condition that  $\beta_2 > 3$  is evidently

$$(k-j)^2(k+j+1) > kj(k+j+2).$$

If  $k = qj$

$$(q-1)^2(q+1 \cdot j+1) > q(q+1 \cdot j+2).$$

Thus, e.g., the curve is leptokurtic if  $q = 4$  when  $j = 2$  so that  $k = 8$ . As an exercise the student may profitably explore the condition that  $\beta_2 = 3$ , and the condition ( $k = 1$ ) that (xvii) defines a monotonic *increasing* function, the mode being then at the upper limit of the range.

*The Type VI distribution.* Types I and II define a distribution of restricted range in virtue of the integral limits of (xvi), i.e. from 0 to  $a$ . We now recall the alternative definition



(p. 251, Vol. I) of the Beta function on the right of (xv), which we may write in a more general form as

$$\frac{k^n}{B(m,n)} \int_0^\infty \frac{x^{m-1} \cdot dx}{(k+x)^{m+n}} = 1 \quad . \quad . \quad . \quad . \quad . \quad (\text{xxi})$$

Accordingly, we may define a Beta variate of unrestricted range in the positive domain as

$$f(x) = \frac{k^n \cdot x^{m-1}}{B(m, n)(k+x)^{m+n}} \quad . \quad . \quad . \quad . \quad . \quad (\text{xxii})$$

This is Pearson's Type VI, the special interest of which emerges from a consideration of its moments. For the  $r$ th zero moment of such a distribution we have by definition

$$\mu_r = \frac{k^n}{B(m, n)} \int_0^\infty \frac{x^{r+m-1}}{(k+x)^{m+n}} \cdot dx.$$

To get the integral in the expression on the right into the same form as (xxi) we may put  $(r + m) = z$ , so that  $m = (z - r)$  and

[illegible]

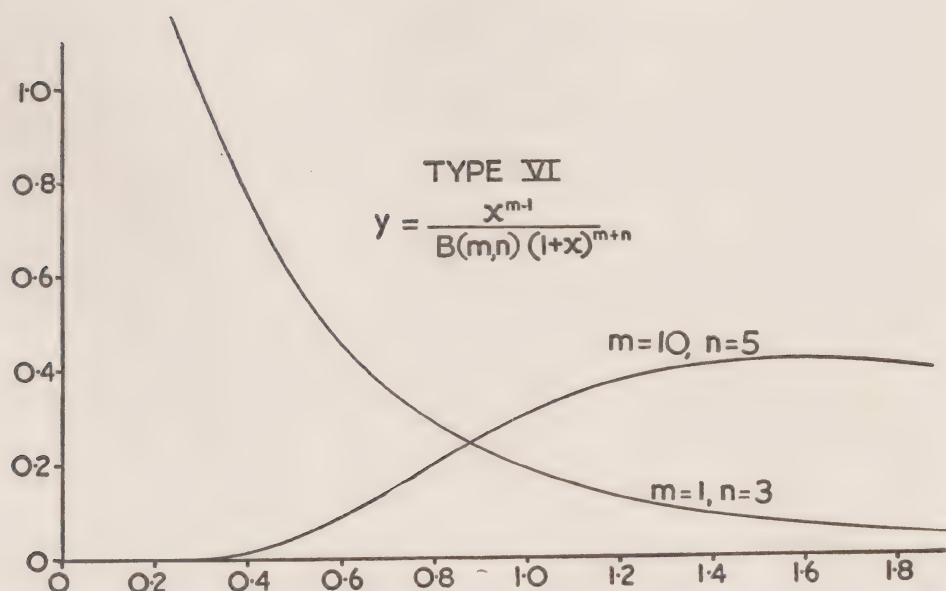


FIG. 115. Pearson's Type VI (Snedecor's  $F$ —see p. 700).





The expression J-shaped in this context is current usage ; and it is not an altogether happy one, since the capital J has an upward bend at its extremity. Like the rectangular hyperbola, the curve under discussion tails off monotonically from a flying start. A J-case of particular interest arises when  $m = \frac{1}{2}$  and  $k = 1$ , so that

$$y = \frac{1}{B(\frac{1}{2}, n)} \cdot \frac{x^{-\frac{1}{2}}}{(1+x)^{n+\frac{1}{2}}} \quad . \quad . \quad . \quad . \quad . \quad (\text{xxv})$$

This we shall see is the parent of Type VII. When the Type VI curve is unimodal in the positive domain, we may explore its shape by evaluating the first four moments. We may disregard the scalar constant by putting  $k = 1$ , in which case we obtain from (xxiii)

$$\begin{aligned} \mu_1 &= \frac{m}{n-1}; & \mu_2 &= \frac{m(m+1)}{(n-1)(n-2)}; \\ \mu_3 &= \frac{m(m+1)(m+2)}{(n-1)(n-2)(n-3)}; & \mu_4 &= \frac{m(m+1)(m+2)(m+3)}{(n-1)(n-2)(n-3)(n-4)}. \end{aligned}$$

Hence we obtain the mean moments in the usual way :

$$m_2 = \frac{m(m+n-1)}{(n-1)^2(n-2)}; \quad m_3 = \frac{2m(2m+n-1)(m+n-1)}{(n-1)^3(n-2)^{(2)}};$$

$$m_4 = \frac{3m(m+n-1)}{(n-1)^4(n-2)^{(3)}}\{m(n+5)(m+n-1) + 2(n-1)^2\}.$$

Thus the first two Pearson coefficients are

$$\beta_1 = \frac{4(n-2)(2m+n-1)^2}{m(n-3)^2(m+n-1)};$$

$$\beta_2 = 3 \left[ \frac{(n+5)(n-2)}{(n-3)(n-4)} + \frac{2(n-1)^2(n-2)}{m(m+n-1)(n-3)(n-4)} \right].$$

For positive values of  $m$  and  $n$  each greater than unity, the distribution, as we have seen it to be, is unimodal. The second mean moment is infinite if  $n = 1$  or  $n = 2$ , the same being true of the third if  $n = 1$  or  $n = 2$  or  $n = 3$  and of the fourth if  $n = 1, n = 2, n = 3$  or  $n = 4$ . For values of  $m$  greater than unity and of  $n$  greater than 4, the value of  $\beta_2$  is greater than 3. For integral positive values of both  $m$  and  $n$  Type VI cannot therefore define a platykurtic distribution.

*Type VII.* The distribution last dealt with is that of the variance ratio of Chapter 16, and as such the basis of Snedecor's  $F$ -test. The  $t$ -distribution commonly called *Student's* in conformity with the pen-name of its author, W. S. Gossett, is Pearson's Type VII. To clarify its genesis, we may remind ourselves of the relation between the normal distribution of the  $c$ -score of unit variance and the distribution of its square ( $C = c^2$ ), i.e. Chi-Square for 1 d.f. We may express this relation by saying that the normal distribution of unit variance is that of the square root of a Chi-Square variate of 1 d.f., and might permissibly speak of the latter as the parent of the normal distribution. The particular form Type VI assumes when  $k = 1$  and  $m = \frac{1}{2}$  in (xxii) is the parent of Type VII in the same sense. The Type VI distribution then defines a variate  $x = z^2$ , such that

$$f(x) = \frac{1}{B(\frac{1}{2}, n)} \cdot \frac{x^{-\frac{1}{2}}}{(1+x)^{n+\frac{1}{2}}}.$$





Whence we obtain

$$m_2 = \frac{1}{2(n-1)}; \quad m_4 = \frac{3}{4(n-1)(n-2)}; \quad m_6 = \frac{15}{8(n-1)(n-2)(n-3)}.$$

The *Student* or *t*-distribution is a particular case of (xxvi) involving a scalar change, *viz.* :

$$z = \frac{t}{\sqrt{f}} \quad \text{and} \quad f = 2n.$$

Whence by substitution in (xxvi)

$$f(t) = \frac{1}{\sqrt{f} \cdot B(\frac{1}{2}, \frac{1}{2}f)} \cdot \frac{1}{\left(1 + \frac{t^2}{f}\right)^{\frac{1}{2}(1+f)}} \quad \dots \quad \text{(xxviii)}$$

For the moments we therefore write

$$m_{2r}(z) = \frac{m_{2r}(t)}{f^r} \quad \text{or} \quad m_{2r}(t) = f^r \cdot m_{2r}(z).$$

Hence from (xxvii)

$$m_{2r}(t) = \frac{f^r \cdot (r - \frac{1}{2})^{(r)}}{(\frac{1}{2}f - 1)^{(r)}}.$$

Whence we obtain

$$m_2 = \frac{f}{f-2}; \quad m_4 = \frac{3f^2}{(f-2)(f-4)}; \quad m_6 = \frac{15f^3}{(f-2)(f-4)(f-6)}.$$

Thus we get

$$\beta_2 = \frac{3(f-2)}{(f-4)} = \frac{3(\overline{f-4} + 2)}{(f-4)} = 3 + \frac{6}{(f-4)} \quad \dots \quad \text{(xxix)}$$

$$\beta_4 = \frac{15(f-2)^2}{(f-4)(f-6)} = 15 + \frac{30(3f-10)}{(f-4)(f-6)} \quad \dots \quad \text{(xxx)}$$

If  $f > 4$  the curve described by (xxviii) is leptokurtic, since  $\beta_2 > 3$ . When  $f = 2$  the variance is infinite.

It is also evident that the fourth mean moment is infinite if either  $f = 2$  or  $f = 4$ . Of more interest is how closely it approaches the normal when  $f$  is large. From the preceding formulae we derive

$f$	$\beta_2$	$\beta_4$	$\beta_6$
16	3.5	24.5	300.1
22	3.3	20.8	208.0
40	3.16	17.7	147.1
60	3.1	16.7	130.3
<i>Normal</i>	3.0	15.0	105.0

At the  $2\sigma$  level the correspondence between the normal and the *t*-distribution is very close when  $f \geq 50$ .





## 15.05 THE TABULATION OF THE GAMMA FUNCTION

The mere fact that the incomplete Gamma and Beta functions are suitable to describe unimodal distributions whose zero moments are always expressible in terms of the complete Gamma function would not justify their prominent role in statistical theory, if it were not also true that they satisfy the practical demands of ready tabulation; and we have hitherto (6.05, Vol. I) indicated the numerical evaluation of  $\Gamma(n)$  only when  $n$  is a positive integer or an odd integer multiple of  $\frac{1}{2}$ , negative or positive. The aim of *Chance and Choice* is to put the playing cards of statistics face upwards on the classroom table; and we should fail to fulfil it if we gave no indication of how it is possible to construct a table of  $\Gamma(n)$  for all values.

What we may call the official procedure relies on the possibility of expressing the Beta function as a trigonometrical integral in accordance with (iii) of 6.07 in Vol. I; but it will suffice for our purpose if we take advantage of the method employed in 6.02 to exhibit the possibility of evaluating the normal integral and of obtaining the standard score which corresponds to the so-called *probable error*. To do this, we made use of the method of integration by series, a trick which is always justifiable if we can express the integrand as a convergent series, and always convenient if the series converges rapidly. It is not easy to fulfil the last condition but the following example will suffice to show that it is possible to evaluate  $\Gamma(n)$  for fractional values of  $n$  other than  $n = \frac{1}{2}$ . We first note that

$$\begin{aligned}
 B(n, n) &= \int_0^1 x^{n-1}(1-x)^{n-1} dx \\
 &= \int_0^1 x^{n-1} (1 - \overline{n-1}_{(1)}x + \overline{n-1}_{(2)}x^2 - \overline{n-1}_{(3)}x^3 \dots) dx \\
 &= \int_0^1 (x^{n-1} - \overline{n-1}_{(1)}x^n + \overline{n-1}_{(2)}x^{n+1} - \overline{n-1}_{(3)}x^{n+2} \dots) dx \\
 &= \left[ \frac{x^n}{n} - \frac{(n-1)_{(1)}x^{n+1}}{1!(n+1)} + \frac{(n-1)_{(2)}x^{n+2}}{2!(n+2)} - \frac{(n-1)_{(3)}x^{n+3}}{3!(n+3)} \dots \right]_0^1 \\
 &= \sum_{r=0}^{\infty} (-1)^r \frac{(n-1)^{(r)}}{r!(n+r)} \dots \dots \dots (i)
 \end{aligned}$$

Let us now suppose that we wish to find the value of  $\Gamma(\frac{3}{4})$  or  $\Gamma(-\frac{1}{4})$

$$B(\frac{3}{4}, \frac{3}{4}) = \frac{\Gamma(\frac{3}{4})\Gamma(\frac{3}{4})}{\Gamma(1\frac{1}{2})}.$$

Since

$$\begin{aligned}
 \Gamma(1\frac{1}{2}) &= \Gamma(\frac{1}{2} + 1) = \frac{1}{2}\Gamma(\frac{1}{2}) = \frac{1}{2}\sqrt{\pi}; \\
 \Gamma(\frac{3}{4}) &= \sqrt{\frac{1}{2}B(\frac{3}{4}, \frac{3}{4})(\pi)^{\frac{1}{2}}} \dots \dots \dots (ii)
 \end{aligned}$$

Likewise

$$\begin{aligned}
 \Gamma(\frac{3}{4}) &= \Gamma(1 - \frac{1}{4}) = -\frac{1}{4}\Gamma(-\frac{1}{4}), \\
 \therefore \Gamma(-\frac{1}{4}) &= -4\sqrt{\frac{1}{2}B(\frac{3}{4}, \frac{3}{4})(\pi)^{\frac{1}{2}}} \dots \dots \dots (iii)
 \end{aligned}$$

In (i) we now insert the value  $n = \frac{3}{4}$ , so that  $(n + r) = \frac{1}{4}(4r + 3)$ , and

$$(n - 1)^{(1)} = -\frac{1}{4}; \quad (n - 1)^{(2)} = -\frac{1}{4}\left(-\frac{1}{4} - 1\right) = \frac{1 \cdot 5}{4 \cdot 4};$$

$$(n - 1)^{(3)} = -\frac{1}{4}\left(-\frac{1}{4} - 1\right)\left(-\frac{1}{4} - 2\right) = \frac{-1 \cdot 5 \cdot 9}{4 \cdot 4 \cdot 4}$$

and in general

$$(n - 1)^{(r)} = (-1)^r \cdot 4^{-r} \prod_{x=1}^{x=r} (4x - 3) \quad \dots \quad (iv)$$

For brevity we may write

$$\begin{aligned} K_r &= \frac{(-1)^r (n - 1)^{(r)}}{r! (n + r)} \\ &= \frac{\prod_{x=1}^{x=r} (4x - 3)}{r! 4^{r-1} (4r + 3)} \quad \dots \quad (v) \end{aligned}$$

Thus (i) becomes

$$B\left(\frac{3}{4}, \frac{3}{4}\right) = \sum_{r=0}^{r=\infty} K_r \quad \dots \quad (vi)$$

The series so defined is convergent. From (v) we obtain

$$\begin{aligned} K_0 &= \frac{4}{3}; \quad K_1 = \frac{1}{7}; \quad K_2 = \frac{1 \cdot 5}{2! \cdot 4 \cdot 11} = \frac{5}{88}; \quad K_3 = \frac{1 \cdot 5 \cdot 9}{3! \cdot 4^2 \cdot 15} = \frac{1}{32}; \\ K_4 &= \frac{1 \cdot 5 \cdot 9 \cdot 13}{4! \cdot 4^3 \cdot 19} = \frac{5 \cdot 8 \cdot 5}{2 \cdot 9 \cdot 184}; \quad K_5 = \frac{1 \cdot 5 \cdot 9 \cdot 13 \cdot 17}{5! \cdot 4^4 \cdot 23} = \frac{6 \cdot 6 \cdot 3}{4 \cdot 7 \cdot 104}. \end{aligned}$$

After the first 5 terms the fall-off is very slow. The first 12 terms add up to 1.64067. If we insert this value and that of  $\sqrt{\pi} \simeq 1.7725$  in (ii) we get  $\Gamma(\frac{3}{4}) \simeq 1.21$  a result correct to 2 significant figures only, the correct value to 5 significant figures being 1.2255. The sum of the first 36 terms yields the value 1.2162 which is correct to 3 significant figures.

#### 15.06 SCORE-SUM AND MEAN SCORE OF A SAMPLE FROM A GAMMA UNIVERSE

We shall now proceed to establish what is the most important property of a Gamma universe in connexion with the theory of sample distributions. By a Gamma universe we imply a unit sample distribution specified by the Gamma variate whose p.d. equation is:

$$y = \frac{k^n \cdot e^{-kx} \cdot x^{n-1}}{\Gamma(n)} \quad \dots \quad (i)$$

The first zero moment and the mean moments are as derived from (vii) in 15.04, viz.:

$$\begin{aligned} \mu_1 &= k^{-1}n & m_4 &= 3k^{-4}n(n + 2) \\ m_2 &= k^{-2}n & m_5 &= 4k^{-5}n(5n + 6) \\ m_3 &= 2k^{-3}n & m_6 &= 5k^{-6}n(3n^2 + 26n + 24). \end{aligned}$$



Whence we have

$$\beta_1 = \frac{4}{n}; \quad \beta_3 = \frac{8(5n+6)}{n^2} \quad . \quad . \quad . \quad . \quad . \quad . \quad (\text{ii})$$

$$\beta_2 = 3 + \frac{6}{n}; \quad \beta_4 = 15 + \frac{10(13n+12)}{n^2} . \quad . \quad . \quad . \quad (\text{iii})$$

We now recall results obtained in 14.05 (viii)–(xi). For the Pearson coefficients of the  $t$ -fold sample score-sum we write accordingly :

$$\begin{aligned}({}_t)\beta_1 &= \frac{1}{t}\beta_1; \quad ({}_t)\beta_2 = 3 + \frac{\beta_2 - 3}{t}; \\({}_t)\beta_3 &= \frac{10(t-1)}{t^2}\beta_1 + \frac{1}{t^2}\beta_3; \\({}_t)\beta_4 &= 15 + \frac{15(t-1)(\beta_2 - 3) + (\beta_4 - 15) + 10(t-1)\beta_1}{t^2}.\end{aligned}$$

By substitution from (ii) and (iii) in the above, we get

$$({}_t)\beta_1 = \frac{4}{nt}; \quad ({}_t)\beta_3 = \frac{8(5nt + 6)}{n^2t^2} \quad \text{(iv)}$$

$${}_{(t)}\beta_2 = 3 + \frac{6}{nt}; \quad {}_{(t)}\beta_4 = 15 + \frac{10(13nt + 12)}{n^2t^2} \quad . \quad . \quad . \quad . \quad (v)$$

In these expressions  $nt$  replaces  $n$  of (ii) and (iii) above. Thus they define the first 4 Pearson coefficients of a unit sample distribution defined by

$$f(S) = \frac{k^{nt} \cdot e^{-kS} \cdot S^{nt-1}}{\Gamma(nt)} \quad (vi)$$

This result suggests the following important rule : *if the unit sample distribution is that of a Gamma variate, we obtain the distribution of the score-sum of the  $t$ -fold sample by substitution of  $(tn)$  for  $n$  in the exponent of the expression definitive of the distribution of the unit sample.* If we put  $k = \frac{1}{2}$  and  $n = \frac{1}{2}f$  in (i), we have  $k = \frac{1}{2}$  and  $nt = \frac{1}{2}ft$  in (vi). We may therefore express the preceding rule by the statement : *if the unit sample distribution of a universe is a Chi-Square variate of  $f$  degrees of freedom, that of the  $t$ -fold sample score-sum (and hence mean) is a Chi-Square variate of  $ft$  degrees of freedom, and hence the sum of  $t$  independent square normal standard scores is a Chi-Square variate of  $t$  degrees of freedom, i.e.*

$$f(S) = \frac{\left(\frac{1}{2}\right)^{\frac{1}{2}t} \cdot e^{-\frac{1}{2}S} \cdot S^{\frac{1}{2}(t-2)}}{\Gamma(\frac{1}{2}t)} . . . . . (\text{vii})$$

The preceding derivation does not constitute a watertight proof of the rule, though it shows that (vi) is likely to give a very good fitting curve for the  $t$ -fold score-sum ( $S$ ) distribution of a sample from the universe of which (i) defines that of the unit sample. We can establish in more than one way the conclusion that all the moments of the  $S$ -distribution and hence also the Pearson coefficients of any order are identical with those of (vi).

We first suppose that  $A$  and  $B$  are Gamma variates with the same scalar constant  $k$ , defined by

$$F(A) = \frac{k^n}{\Gamma(n)} e^{-kA} A^{n-1} \text{ and } F(B) = \frac{k^m}{\Gamma(m)} e^{-kB} B^{m-1},$$

$$\therefore \mu_r(A) = k^{-r}(n+r-1)^{(r)} \text{ and } \mu_r(B) = k^{-r}(m+r-1)^{(r)}.$$

For the moments of the score-sum distribution of independent samples whose distributions conform to the above we have

$$\begin{aligned} \mu_r(A+B) &= E(A+B)^r = \sum_{x=0}^{x=r} r_{(x)} E(A^x) \cdot E(B^{r-x}), \\ \therefore \mu_r(A+B) &= \sum_{x=0}^{x=r} r_{(x)} \cdot \mu_x(A) \cdot \mu_{r-x}(B) \\ &= k^{-r} \sum_{x=0}^{x=r} r_{(x)} (n+x-1)^{(x)} (m+r-x-1)^{(r-x)}, \\ \therefore \mu_r(A+B) &= k^{-r} \cdot r! \sum_{x=0}^{x=r} \frac{(n+x-1)^{(x)}}{x!} \frac{(m+r-x-1)^{(r-x)}}{(r-x)!}. \end{aligned}$$

In the notation of Chapter 1 of Vol. I this expression involves a product of Figurates, *viz.*:

$$\mu_r(A+B) = k^{-r} \cdot r! \sum_{x=0}^{x=r} {}^x F_n \cdot {}^{r-x} F_m.$$

Whence from (x) in 11.07

$$\mu_r(A+B) = k^{-r} \cdot r! {}^r F_{n+m} = k^{-r}(n+m+r-1)^{(r)}.$$

The above expression defines the moments of the distribution of a score  $C(=A+B)$  defined by the Gamma variate

$$F(C) = \frac{k^{n+m}}{\Gamma(n+m)} e^{-kC} C^{n+m-1}.$$

Thus the score-sum of a  $\Gamma(n)$  and of a  $\Gamma(m)$  variate is a  $\Gamma(n+m)$  variate. Hence that of two  $\Gamma(n)$  variates as defined by (i) is that of a  $\Gamma(2n)$  variate, that of three  $\Gamma(n)$  variates is a  $\Gamma(3n)$  variate and so on.

*Alternatively*, we may reach the same result by recourse to the moment generating function of the Gamma variate. If  $f(x)$  is the p.d. of a variate  $x$  whose range is from 0 to  $\infty$ , the m.g.f. of the distribution is given by

$$G(\mu) = \int_0^\infty e^{xt} f(x) \cdot dx.$$

For the m.g.f. of the unit sample distribution defined by (i) above, we therefore have

$$\begin{aligned} G(\mu) &= \frac{k^n}{\Gamma(n)} \int_0^\infty e^{-(k-t)x} x^{n-1} dx \\ &= \frac{k^n}{\Gamma(n)} \cdot \frac{\Gamma(n)}{(k-t)^n}, \\ \therefore G(\mu) &= k^n \cdot (k-t)^{-n} = \left(1 - \frac{t}{k}\right)^{-n}. \end{aligned}$$



The m.g.f. of the  $s$ -fold score-sum sample distribution in accordance with (iii) of 14.02 is therefore

$$G_s(\mu) = \left(1 - \frac{t}{k}\right)^{-sn}.$$

We obtain the general expression for the moments of the distribution of the score-sum of the  $s$ -fold sample as in 14.04 by recourse to the relation

$$\mu_r = \frac{d^r}{dt^r} G_s(\mu) \text{ when } t = 0.$$

By successive differentiation we have

$$\begin{aligned} \frac{d^r}{dt^r} G_s(\mu) &= (sn + r - 1)^{(r)} k^{sn} (k - t)^{-sn-r}, \\ \therefore \mu_r &= k^{-r} (sn + r - 1)^{(r)}. \end{aligned}$$

This defines the moments of the distribution in which  $sn$  replaces  $n$  of (i) above. Thus the score-sum of an  $s$ -fold sample from a  $\Gamma(n)$  universe is a  $\Gamma(sn)$  variate.

*The so-called Chi-Square Test.* It will forestall misunderstanding at a later stage if we take this opportunity to comment on the use of the expression *Chi-Square test*, recalling earlier remarks on the  $c$ -test. In different chapters of Vol. I we have referred to a  $c$ -test with judicious use of the indefinite article. A  $c$ -test is a test we can rightly apply to a score whose distribution is approximately normal; but whether a score distribution approximates to the normal form and with what order of precision is a matter for separate enquiry with due regard to the nature of the score. The normal curve is what the Herbals describe as a protean genus which turns up in unlikely localities; and the rationale for invoking it in one context, e.g. quality control, has no intrinsic relation to the reason for enlisting its aid in another, e.g. the proportionate score difference of two sub-samples.

Similar remarks apply with equal force to what many current treatises refer to as the *Chi-Square* ( $\chi^2$ ) test. The expression recalls a well-known comment on the Lord Privy Seal, a personage who is not necessarily a lord, never a privy and in no sense a seal. A species of the Pearson Type III genus turns up as a sampling distribution in diverse situations for reasons just as diverse; and tables of the corresponding integral are in use to test many different hypotheses. The statistic commonly referred to as Chi-Square is in fact a sum of squares, and is not itself a square, except when we speak as below of Chi-Square for 1 degree of freedom.

\* \* \* \* \*

*Pooling Data.* The additive property of the Chi-Square variate set forth in this section calls for a caveat w.r.t. a recipe sometimes, and in the opinion of the writer wrongly, cited for situations in which the end in view is to assess the significance of an assemblage of normal tests the result of none of which is highly significant *per se*. We have seen that a normal test, i.e. a  $c$ -test, is on all fours with a Chi-Square test for 1 d.f. if we use the table of the latter to assess the probability of getting a score as great as  $c^2 = C$ ; but we have to take something for granted when we do this. By squaring  $c$  we eliminate the sign difference. Hence assessing the probability that the sum of a set of  $c$ -scores (e.g.  $d_1 = 1.5\sigma_1$ ,  $d_2 = 2.1\sigma_2$ ,  $d_3 = 1.9\sigma_3$ , etc.), each referable to a difference in the *same direction* though each or most of them numerically too low to inspire confidence, will attain its particular numerical value is not on all fours with assessing the probability that a sum of Chi-Square variates will have a particular numerical value.

In such a situation, elementary considerations may supply the answer we seek. For instance, we may suppose that we perform six experiments or make six sets of observations involving a difference, e.g. percentage of persons not attacked in an epidemic when treated in one or other





2. In the same way analyse the following data compiled by Greenwood, w.r.t. efficacy of prophylactic inoculation, from official returns of the Western Front 1914-18.

Year	Disease	Incidence per 1000		Death-rate per 1000		Case Mortality per 1000		Number of Cases	
		Pro- tected	Unpro- tected	Pro- tected	Unpro- tected	Pro- tected	Unpro- tected	Pro- tected	Unpro- tected
1914	Typhoid	—	—	—	—	5.8	17.3	51	202
	Para A	—	—	—	—	—	—	—	5
	Para B	—	—	—	—	—	3.2	—	31
1915	Typhoid	0.93	8.1	0.007	1.8	7.5	23.2	517	288
	Para A	—	0.4	—	0.003	—	0.7	—	281
	Para B	—	1.7	—	0.03	—	1.9	—	1043
1916	Typhoid	0.57	0.51	0.009	0.04	1.58	8.33	693	36
	Para A	0.21	3.19	0.003	0.05	1.56	1.78	256	224
	Para B	0.3	9.2	0.002	0.07	0.82	0.77	362	647
1917	Typhoid	0.104	1.09	0.008	0.13	7.7	12.12	194	33
	Para A	0.07	1.12	0.000	0.03	—	2.93	139	34
	Para B	0.18	4.14	0.003	0.13	1.7	3.20	346	125
1918	Typhoid	0.02	0.19	0.003	0.04	13.84	24.00	65	25
	Para A	0.01	0.04	0.000	—	2.7	—	37	6
	Para B	0.05	0.22	0.000	—	0.78	—	127	29

3. From the same source we obtain the following figures for effects of inoculation against enteric fever of a regiment (17th Lancers) in Meerut :

	1907	1909
No. inoculated . . . . .	430	460
No. untreated . . . . .	220	160
Total No. . . . .	650	620
Cases among inoculated . . . . .	13	18
<i>ditto</i> untreated . . . . .	95	96
Deaths among inoculated . . . . .	1	2
<i>ditto</i> untreated . . . . .	13	18

### 15.07 THE INDEPENDENCE CONDITION

We have assumed that an  $f$ -fold sample from a universe in the context of the foregoing theorem signifies the same thing as  $f$  independently selected score values. Our conclusion is therefore that the score-sum of  $f$  independent Chi-Square variates of 1 d.f. is a Chi-Square variate of  $f$  degrees of freedom. A question fundamental to the rationale of the significance tests we shall later examine is whether we can assume that  $f$  Chi-Square variates of 1 d.f. are independent if their score-sum is a Chi-Square variate of  $f$  degrees of freedom. When we say that R. A. Fisher gave the first rigorous proof of the so-called *Student* distribution, this is the pivotal issue. For Gosset implicitly assumed that two variates of zero covariance are independent. We have seen that this is not so, though the converse is true, that independence implies zero covariance. To

say that two variates  $a$  and  $b$  are strictly independent implies that the covariance of any integral power of  $a$  and any integral power of  $b$  is zero, i.e. for *all* whole number values of  $n$  and  $m$

$$\text{Cov}(a^n, b^m) = 0.$$

Our models of 12.07 have shown that  $\text{Cov}(a, b) = 0$  does not necessarily imply that this is so; but common sense suffices to justify the conclusion that  $\text{Cov}(a^n, b^m)$  of 11.02, being an index of whether high values of  $a$  more often than otherwise correspond with high or low values of  $b$ , must *have the same sign* regardless of the numerical value of  $n$  or  $m$ , unless zero. Let us therefore examine the implications of deriving the result obtained in 15.04 by the method of moments without assuming independence. If we do assume independence we write

$$E(a^x \cdot b^{k-x}) = E(a^x) \cdot E(b^{k-x}) = \mu_x(a) \cdot \mu_{k-x}(b).$$

Otherwise we must put

$$\begin{aligned} E(a^x \cdot b^{k-x}) &= \text{Cov}(a^x, b^{k-x}) + E(a^x) \cdot E(b^{k-x}) \\ &= \text{Cov}(a^x, b^{k-x}) + \mu_x(a) \cdot \mu_{k-x}(b). \end{aligned}$$

If we do not assume their independence we must therefore write the moments of the distribution of the score-sum of two variates  $a$  and  $b$  in the form:

$$\begin{aligned} \mu_k(a+b) &= E(a+b)^k \\ &= \sum_{x=0}^{x=k} k_{(x)} E(a^x b^{k-x}) \\ &= \sum_{x=0}^{x=k} k_{(x)} \mu_x(a) \cdot \mu_{k-x}(b) + \sum_{x=0}^{x=k} k_{(x)} \text{Cov}(a^x, b^{k-x}). \end{aligned}$$

Hence, if  $\text{Cov}(a^x, b^{k-x}) > 0$  for some value of  $x$ , there must be some value of  $k$  such that

$$\mu_k(a+b) > \sum_{x=0}^{x=k} k_{(x)} \mu_x(a) \cdot \mu_{k-x}(b).$$

Similarly, if  $\text{Cov}(a^x, b^{k-x}) < 0$  for some value of  $x$ , there must be some value of  $k$  such that

$$\mu_k(a+b) < \sum_{x=0}^{x=k} k_{(x)} \mu_x(a) \cdot \mu_{k-x}(b).$$

Hence it must always be true that  $\text{Cov}(a^x, b^{k-x}) = 0$  if

$$\mu_k(a+b) = \sum_{x=0}^{x=k} k_{(x)} \cdot \mu_x(a) \cdot \mu_{k-x}(b).$$

If  $a$  and  $b$  are Chi-Square variates of  $m$  and  $n$  degrees of freedom, the expression on the right defines the moments of a Chi-Square distribution of  $(m+n)$  degrees of freedom. Hence the distribution of the sum of Chi-Square variates of  $m$  and  $n$  degrees of freedom respectively will be a Chi-Square variate of  $(m+n)$  degrees of freedom, if, and only if,  $a$  and  $b$  are independent, and we may extend this conclusion by iteration to the sum of any number of Chi-Square variates.



The series involved converges rapidly, and if we take the first 3 terms only we get  $P_a = 0.4385$ , so that  $(1 - P_a)$  in (ii) has the value 0.5615. On taking 4 terms we get  $P_a = 0.4421$  and





## SIGNIFICANCE TESTS FOR ANALYSIS OF VARIANCE

## 16.01 THE VARIANCE PROBLEM

IN Chapter 13 we explored the possible breakdown of a single sample w.r.t. particular criteria of classification into subsamples with a view to deciding whether variation *inter se* is consistent with the possibility of each being a set from one and the same universe. Our enquiry led us to formulate different estimates of the variance of the score distribution in the putative common universe; and currently prescribed assessment of the credentials of the null hypothesis depends on the consistency of such estimates. So far we have merely asked the question: what statistics of such a set-up must be consistent? We have now to define the criteria of consistency in such a context more explicitly.

To say that two estimates are consistent in the most literal sense of the term means that their difference is zero or that their ratio is unity. When a null hypothesis postulates consistency of two estimates, the hitherto customary test procedure deems agreement to be satisfactory, if the deviation of the observed difference or the observed ratio from its mean value prescribed by the hypothesis is not excessive. This presupposes that we can define the distribution of one or the other. That of the ratio has an advantage by no means obvious at first sight. Since the class of significance tests we are about to examine relies on the distribution of ratios, a brief digression to explain it will be necessary.

Let us suppose that  $A$  and  $B$  are two variates whose distribution involves an *unknown* parameter ( $K$ ), e.g. the true variance in the case of a normal universe of which our only knowledge comes from two samples. It may then be possible to specify exactly the distribution of  $K.A$  and  $K.B$ . If the distribution of the difference ( $A - B$ ) is expressible in terms of the distributions of  $A$  and  $B$ , we can of course define the distribution of  $(KA - KB) = K(A - B)$ ; but we cannot define the distribution of  $(A - B)$  in numerical terms unless we know the value of  $K$ . On the other hand, it may happen that we can derive the distribution of the ratio of  $K.A$  to  $K.B$ ; and if so, we can define the distribution of the ratio  $(A \div B) = (K.A \div K.B)$ . Thus the derivation of the explicit distribution of a ratio may be possible in the absence of information necessary for defining the precise distribution of a difference.

Such is the situation we have to deal with when our concern is the consistency of statistics involving sums of normally distributed square scores. We have elsewhere (14.02, 15.02) seen that we can define the distribution of the sum of normally distributed square scores of zero mean and *unit variance*, i.e. the ratio of square score deviations from an unknown true mean to an unknown true variance of the parent universe. For the reason stated, the fact that we do not know the true variance of the universe distribution gives rise to no difficulty if the distribution of their ratio rather than the distribution of their difference is what concerns us. The fact that we do not need to know the true mean will appear from considerations which follow.

Suppose  $x_r$  is any unit score from a normal universe, that its expected value is  $M$  and that the variance of the unit sample distribution is  $\sigma^2$ . We then define a square standard score as

$$c_r^2 = \frac{(x_r - M)^2}{\sigma^2} \quad . \quad . \quad . \quad . \quad . \quad . \quad (i)$$

In this expression the first term on the right is a Chi-Square variate of  $n$  degrees of freedom. Since the mean score of a random sample from a normal universe is also a normal variate, the second term is a square normal score of unit variance and is therefore a Chi-Square variate of 1 d.f. Thus we have expressed in (vi) the sum of  $n$  square deviations from the *sample mean*



as the difference between two Chi-Square variates. Thus the fact that we do not know the value of the true mean  $M$  need not trouble us, if we can formally define the distribution of the difference of two such Chi-Square variates. To do so, it is customary to rely on mathematical techniques which involve an understanding of matrix algebra and the manipulation of multiple integrals ; but it is possible to exhibit the argument at a more elementary level, if our approach is heuristic. We first explore the result of expressing (vi) in terms of normal scores of unit variance such as (i) above. We then put

$$c_r = \frac{x_r - M}{\sigma} \quad \text{and} \quad x_r = M + \sigma \cdot c_r,$$

$$\therefore n.M_x = \sum_{r=1}^{r=n} x_r = nM + \sum_{r=1}^{r=n} \sigma.c_r,$$

$$\therefore n(M_x - M) = \sigma \sum_{r=1}^{r=n} c_r,$$

$$\therefore \frac{n(M_x - M)^2}{\sigma^2} = \left[ \sum_{r=1}^{r=n} \frac{c_r}{\sqrt{n}} \right]^2.$$

Hence we may write (v) in the form

[illegible]

The two Chi-Square variates on the right of (vi) and (vii) are not independent, since the mean ( $M_x$ ) and the sum of squares are necessarily correlated; but it will be possible to express the statistic on the left as a Chi-Square variate of  $(n - 1)$  degrees of freedom, if we can define a set of  $n$  independent normal scores ( $u_1, u_2 \dots u_n$ ) of zero mean and unit variance such that

$$u_1 = \sum_{r=1}^{r=n} \frac{c_r}{\sqrt{n}} \quad . \quad . \quad . \quad . \quad . \quad (\text{viii})$$

$$u_1^2 + u_2^2 + \dots + u_n^2 = c_1^2 + c_2^2 + \dots + c_n^2 \quad . \quad . \quad . \quad . \quad (\text{ix})$$

$$\therefore \frac{n \cdot V_s}{\sigma^2} = (u_1^2 + u_2^2 + u_3^2 \dots u_n^2) - u_1^2 = \sum_{r=2}^{r=n} u_r^2 \quad . \quad . \quad . \quad (x)$$

From (ix) and (ii), we see that

$$\sum_{r=1}^{r=n} c_r^2 = S = \sum_{r=1}^{r=n} u_r^2.$$

The sum of the squares of all the  $u$ -scores is thus a Chi-Square variate of  $n$  degrees of freedom, *i.e.* that of the sum of  $n$  independent Chi-Square variates of 1 d.f. ; and each square  $u$ -score of the sum on the right of the above is by definition a Chi-Square variate of 1 d.f. Now we have seen (15.07) that the sum of  $n$  Chi-Square variates of 1 d.f. is a Chi-Square variate of  $n$  d.f. only if the former are independent. Hence the sum of  $(n - 1)$  square  $u$ -scores on the right of (x) is a Chi-Square of  $(n - 1)$  degrees of freedom, if the assumption implicit in (viii) and (ix) is admissible, *i.e.* that (viii) and (ix) are consistent with the postulate that each  $u$ -score is a normal score of unit variance. We get a clue to the justification of the postulate if we write (viii) in full as :

$$u_1 = n^{-\frac{1}{2}} \cdot c_1 + n^{-\frac{1}{2}} \cdot c_2 + n^{-\frac{1}{2}} \cdot c_3 \dots n^{-\frac{1}{2}} \cdot c_n \quad , \quad , \quad , \quad (xi)$$

Let us consider the pattern

$$u_r = A_r \cdot c_1 + B_r \cdot c_2 \dots N_r \cdot c_r \quad . \quad . \quad . \quad . \quad (xii)$$

In the above, our  $c$ -scores are of unit variance and are independent. The variance of the distribution of the score  $u_r$  thus follows from (iii) in 14.01, *viz.* :

$$V(u_r) = A_r^2 + B_r^2 + \dots N_r^2.$$

Hence  $u_r$  is a score of unit variance if

$$A_r^2 + B_r^2 \dots + N_r^2 = 1 \quad . \quad . \quad . \quad . \quad (xiii)$$

If  $c_r$  is a normal score of unit variance,  $A_r \cdot c_r$  is a normal score of variance  $A_r^2$ . Hence  $u_r$  is the sum of  $n$  independent normal scores and is therefore itself a normal variate. Thus each of our  $u$ -scores is a normal score of unit variance if (xiii) holds good.

We have now only to show that the condition defined by (ix) is consistent with (xiii). The next step will be easier to generalise, if we illustrate it by the case of  $n = 3$ , so that

$$u_1 = A_1 \cdot c_1 + B_1 \cdot c_2 + C_1 \cdot c_3;$$

$$u_2 = A_2 \cdot c_1 + B_2 \cdot c_2 + C_2 \cdot c_3;$$

$$u_3 = A_3 \cdot c_1 + B_3 \cdot c_2 + C_3 \cdot c_3.$$

In this case

$$\begin{aligned} u_1^2 + u_2^2 + u_3^2 &= (A_1^2 + A_2^2 + A_3^2)c_1^2 + (B_1^2 + B_2^2 + B_3^2)c_2^2 + (C_1^2 + C_2^2 + C_3^2)c_3^2 \\ &\quad + 2(A_1B_1 + A_2B_2 + A_3B_3)c_1c_2 + 2(A_1C_1 + A_2C_2 + A_3C_3)c_1c_3 \\ &\quad + 2(B_1C_1 + B_2C_2 + B_3C_3)c_2c_3. \end{aligned}$$

To ensure that (ix) holds good we must simultaneously make

$$(A_1^2 + A_2^2 + A_3^2) = (B_1^2 + B_2^2 + B_3^2) = (C_1^2 + C_2^2 + C_3^2) = 1 \quad . \quad . \quad (xiv)$$

$$(A_1B_1 + A_2B_2 + A_3B_3) = (A_1C_1 + A_2C_2 + A_3C_3) = (B_1C_1 + B_2C_2 + B_3C_3) = 0 \quad . \quad (xv)$$

In accordance with (xiii) to ensure that our  $u$ -scores are normal scores of unit variance, we must also define them so that

$$(A_1^2 + B_1^2 + C_1^2) = (A_2^2 + B_2^2 + C_2^2) = (A_3^2 + B_3^2 + C_3^2) = 1 \quad . \quad . \quad (xvi)$$

In accordance with (viii), we have already defined

$$A_1^2 = B_1^2 = C_1^2 = \frac{1}{n} \quad . \quad . \quad . \quad . \quad . \quad . \quad (xvii)$$

Thus three conditions only are sufficient to ensure that an  $n$ -fold set of  $u$ -scores defined by (xii) simultaneously satisfy (viii) and (ix), hence also (xii) :

- (a) each of the constants of one of the  $u$ -scores defined by (x) must be equal to  $n^{-\frac{1}{2}}$ ;
- (b) the sum of the squares of the constants in each row ( $A_r^2, B_r^2$ , etc.) and in each column ( $A_1^2, A_2^2 \dots A_n^2, B_1^2, B_2^2 \dots B_n^2$ , etc.) must alike be unity;
- (c) the total sum of cross products of corresponding constants ( $A_iB_i$ , etc.) in any two columns must be zero.

If we can choose the constants  $A_r, B_r$ , etc. in (xii) to satisfy these three conditions simultaneously, we can say that the  $n$ -fold sum of the square  $u$ -scores is equal to the  $n$ -fold sum of the square  $c$ -scores, i.e. that its distribution is that of a Chi-Square variate of  $n$  degrees of freedom ;



and if it is admissible that the sum of  $n$  Chi-Square variates of 1 d.f. is a Chi-Square variate of  $n$  degrees of freedom only if they are independent, we must conclude that our  $u$ -scores are independent variates. If so, the distribution of  $n - 1$  of our  $u$ -scores is a Chi-Square variate of  $(n - 1)$  degrees of freedom, and the statistic defined by the left of (x) is itself a Chi-Square variate of  $(n - 1)$  degrees of freedom.

What we have still to ask therefore is whether we can indeed choose  $A_r$ ,  $B_r$ , etc. to satisfy (a)–(c) simultaneously. The student who is familiar with the elementary theory of equations will realise that we have indeed at our disposal the requisite number of conditions to fix the constants other than  $A_1, B_1$ , etc.  $= n^{-\frac{1}{2}}$ . Others may more easily grasp that this is so, if we illustrate the possibility of satisfying the prescribed conditions by numerical examples as in 16.02 below. What numerical values of the constants other than  $A_1, B_1, C_1 \dots N_1$  satisfy the conditions prescribed are, of course, immaterial to our purpose except to illustrate that a solution is possible. In choosing them, it is important to remember that no constant can be numerically greater than unity; but the sign need not be positive. Indeed, no solution would be possible on that understanding. This is consistent with our statistical requirements because the difference between two normal variates is a normal variate whose variance is equal to that of their sum, i.e.  $(A_{rc_1} - B_{rc_1})$  is a normal variate of variance  $(A_r^2 + B_r^2)$ .

Before proceeding we may make explicit the outcome of the foregoing reasoning. Our concern has been to define the distribution of the  $n$ -fold sum ( $Q$ ) of *standardised* square deviations from the *sample mean*, i.e.

$$\frac{n \cdot V_s}{\sigma^2} = Q = \sum_{r=1}^{r=n} \frac{(x_r - M_x)^2}{\sigma^2} . . . . . (\text{xviii})$$

So defined  $Q$  is a Chi-Square variate of  $f = (n - 1)$  degrees of freedom, i.e.

$$F(Q) = \frac{(\frac{1}{2})^{\frac{1}{2}} f e^{-\frac{1}{2}Q} Q^{\frac{1}{2}(f-2)}}{\Gamma(\frac{1}{2}f)} \quad . \quad . \quad . \quad . \quad . \quad (\text{xix})$$

The statistic whose mean value is the true variance of the score distribution in the parent universe is as defined in 13.02 :

$$\frac{1}{(n-1)} \sum_{r=1}^{r=n} (x_r - M_w)^2 = s^2 = \frac{n}{f} V_s,$$

$$\therefore \frac{s^2}{\sigma^2} = \frac{Q}{f} = S_c.$$

We may speak of this statistic as the unbiased estimate of the variance in *standard form*. By definition, therefore,

$$\frac{dQ}{dS_c} = f.$$

Whence we may write in accordance with Case I of 15.02

$$\phi(S_c) = \frac{\left(\frac{f}{2}\right)^{\frac{1}{2}f} . e^{-\frac{1}{2}f.S_c} . S_c^{\frac{1}{2}(f-2)}}{\Gamma(\frac{1}{2}f)} . . . . . (\text{xx})$$

The last equation has the same form as (viii) in 15.05; but  $f = (n - 1)$  replaces  $n$  in the latter which describes the distribution of the mean value of the sum of  $n$  square deviations of unit variance from the true mean.

## EXERCISE 16.01

*Note.*—The following score transformations are not orthogonal, but may help the student to materialise some implications of such a change from one score system to another.

1. For the Lottery Model of 11.08 (Fig. 91) evaluate the sampling distribution of  $u_a = x_1 + 3x_2$ ;  $u_b = 3x_1 - x_2$ .

2. If  $x_1$  and  $x_2$  are the results of the first and second toss of the double spin of a coin, determine the sampling distribution of  $S_2 = u_a^2 + u_b^2$ , if  $u_a$  and  $u_b$  have the same meaning as in 1.

3. For the 3-fold spin of a tetrahedral die with faces having 1, 2, 2, 3 pips, the first, second and third unit scores being  $x_1, x_2, x_3$ , we may define a score system

$$u_a = x_1 + 3x_2 + 2x_3; \quad u_b = 3x_1 - x_2 + 2x_3; \quad u_c = 2x_1 - x_2 - x_3.$$

Determine the distribution of the sample score  $S_3 = u_a^2 - u_b^2 + 2u_c^2$ .

## 16.02 THE ORTHOGONAL TRANSFORMATION

Without invoking any considerations other than those dictated by the statistical requirements of the problem, we have developed in 16.01 a score transformation suggested by the properties of the *Orthogonal Lottery Model* (Fig. 97). The rationale of many significance tests devised during the last three decades invokes such score transformations the practicability and numerical meaning of which are easily demonstrable without recourse to higher mathematics. One step in the argument calls for clarification. The reader who is not familiar with the theory of equations will want assurance. We shall need  $n^2$  constants  $A_r, B_r$ , etc. to satisfy (ix); and of these  $n$  must have the value  $A_r = n^{-\frac{1}{2}} = B_r$ , etc. to satisfy (viii) and (xiii). Can we choose the remaining  $n(n-1)$  constants to satisfy (xiv)–(xvi)? Such is the theme of this section; but it may be helpful to some readers who have unsuccessfully tackled a more advanced treatment if we first clarify the historical background of the test procedures dealt with below.

The customary approach to the score transformation outlined in 16.01 is intelligible in the context of R. A. Fisher's earliest work, published when the impact of the theory of relativity on experimental physics had lately provoked interest in the abstract geometry of the hypersphere. In this setting, there were new clues for the mathematician and new difficulties for the practical statistician unfamiliar with the new geometries. For at least two decades very few among those who espoused the techniques of Fisher were indeed equipped to evaluate their mathematical credentials. Fortunately, familiarity with the mathematical tools which Fisher's school relied on is not essential to an understanding of the outcome, and the reader who skips the ensuing brief digression will not be at a disadvantage.\*

\*\*\*\**The Geometrical Analogy.* The pattern for the transformation of the 2-fold sample of unit scores is

$$\begin{aligned} y_1 &= a_{11} \cdot x_1 + a_{12} \cdot x_2 \quad \text{and} \quad x_1^2 + x_2^2 = y_1^2 + y_2^2. \\ y_2 &= a_{21} \cdot x_1 + a_{22} \cdot x_2. \end{aligned}$$

This will be a familiar lay-out to the student who has gone far in co-ordinate geometry, being the usual jumping-off ground for an introduction to matrix algebra; but its interpretation should in any case offer no difficulty. Let us suppose that  $P$  is a point in a plane whose co-ordinates with respect to one Cartesian grid are  $x_1$  and  $x_2$ . If  $r$  is its distance (the *radius vector*

\* Four stars mark both the beginning and the end of the passage referred to.



of the point) from the origin of the grid, the theorem of Pythagoras prescribes that  $r^2 = x_1^2 + x_2^2$ . Let us now visualise the same point in a second grid with the same origin and hence the same radius vector. If its co-ordinates in the latter are  $y_1$  and  $y_2$ , we may also write  $r^2 = y_1^2 + y_2^2$ . Hence  $x_1^2 + x_2^2 = y_1^2 + y_2^2$ . The condition  $(x_1^2 + x_2^2) = (y_1^2 + y_2^2)$  thus suffices to justify the geometrical interpretation of two linear equations involving 2 independent variables as a *rotation of axes*. If  $a$  is the angle the second grid makes with the first, elementary trigonometry (Fig. 117) suffices to show that

$$y_1 = \cos a \cdot x_1 - \sin a \cdot x_2;$$

$$y_2 = \sin a \cdot x_1 + \cos a \cdot x_2.$$

If we now write  $a_{11} = \cos a = -a_{22}$  and  $a_{12} = \sin a = a_{21}$ , the following identities follow :

$$a_{11}^2 + a_{12}^2 = 1 = a_{21}^2 + a_{22}^2 \text{ and } a_{11}^2 + a_{12}^2 = 1 = a_{12}^2 + a_{22}^2 \quad . \quad . \quad . \quad (i)$$

(since  $\sin^2 a + \cos^2 a = 1$ ).

$$a_{11} \cdot a_{12} + a_{21} \cdot a_{22} = 0 = a_{11} \cdot a_{21} + a_{12} \cdot a_{22} \quad . \quad . \quad . \quad (ii)$$

### ROTATION OF AXES IN A PLANE

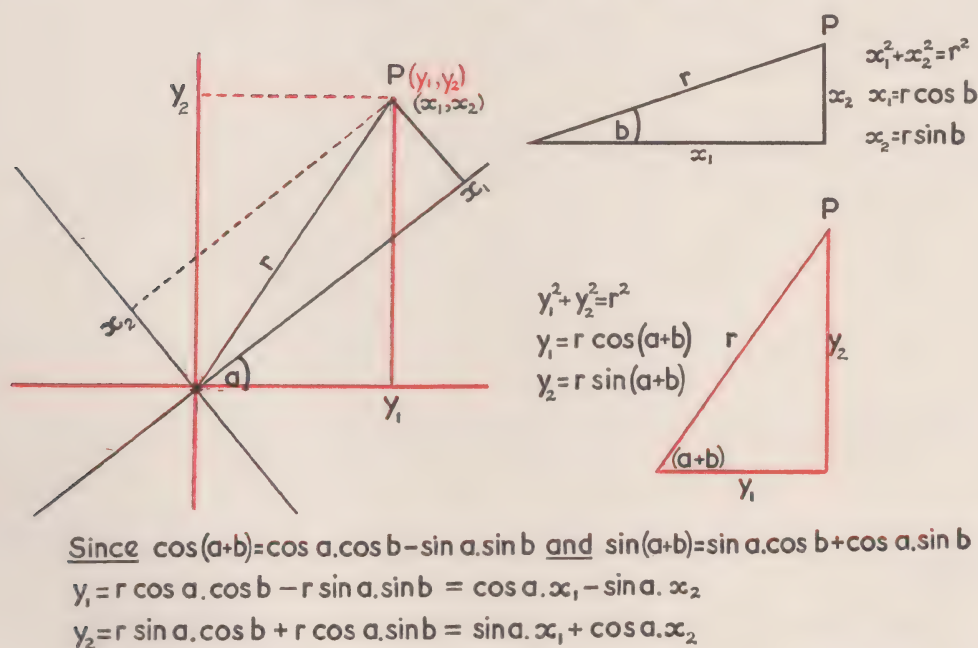


FIG. 117. Geometrical meaning of the Orthogonal Transformation.

$$\text{If } a = 45^\circ, \cos a = \frac{1}{\sqrt{2}} = \sin a \text{ and } y_1 = \frac{1}{\sqrt{2}}x_1 + \frac{1}{\sqrt{2}}x_2; \quad y_2 = \frac{1}{\sqrt{2}}x_1 - \frac{1}{\sqrt{2}}x_2.$$

If we interpret our score transformation in terms of the rotation of axes, we thus see that :  
 (a) each horizontal and each vertical sum of the squares of the coefficients is equal to unity ;  
 (b) the sum of all the vertical cross products and of all the horizontal cross products is zero.

The reader who has an elementary knowledge of co-ordinate geometry in 3-dimensions will be able to take the argument a step further. If  $P$  is a point in 3-dimensional space with co-ordinates  $x_1, x_2, x_3$ , the equation which defines its radius vector ( $r$ ) is  $x_1^2 + x_2^2 + x_3^2 = r^2$ . If, therefore,  $y_1, y_2, y_3$  are its co-ordinates in another framework with the same origin,

$$y_1^2 + y_2^2 + y_3^2 = r^2 = x_1^2 + x_2^2 + x_3^2.$$

We may then specify the relation between the co-ordinates in the form :

$$y_1 = a_{11} \cdot x_1 + a_{12} \cdot x_2 + a_{13} \cdot x_3 ;$$

$$y_2 = a_{21} \cdot x_1 + a_{22} \cdot x_2 + a_{23} \cdot x_3 ;$$

$$y_3 = a_{31} \cdot x_1 + a_{32} \cdot x_2 + a_{33} \cdot x_3.$$

By elementary trigonometry, we may derive relations similar to those of (i) and (ii) above.

In 11.07 we have metaphorically spoken of *super-solid* figurate numbers, the build-up of which follows the same lines as the corresponding picturable number scores for 0, 1, 2 and 3 dimensions. In the same way, there is no objection to the use of the term radius vector to define a sum of squares such as  $r^2 = (x_1^2 + x_2^2 + x_3^2 + x_4^2)$ , and we may speak metaphorically of  $x_1, x_2$ , etc. as co-ordinates of a point in a 4-dimensional ultra-visual grid. If  $r^2 = (y_1^2 + y_2^2 + y_3^2 + y_4^2)$  we may likewise speak of  $y_1, y_2$ , etc. as co-ordinates of the same point in a 4-dimensional grid with the same origin, and define a 4-fold system of linear equations descriptive of rotation of axes in a so-called 4-dimensional space. Actually, we are here using the idiom of geometry to describe algebraic manipulations which we can no longer picture. Whether it is helpful or otherwise to do so depends entirely on whether it is a familiar idiom. If so, an excursion into the hypersphere may help us to discover new or to interpret known relations. Otherwise, the safe course is to keep our feet on the solid ground of algebra, as we shall now do. In any case, we have to rely on matrix algebra to generalise the rules which we shall now examine in greater detail.

*Rules of the Transformation.* As a general pattern of the so-called orthogonal transformation, it will suffice to lay out a 4-fold set of duplicate scores  $u_r$  and  $x_s$  :

$$u_1 = A_1x_a + B_1x_b + C_1x_c + D_1x_d ;$$

$$u_2 = A_2x_a + B_2x_b + C_2x_c + D_2x_d ;$$

$$u_3 = A_3x_a + B_3x_b + C_3x_c + D_3x_d ;$$

$$u_4 = A_4x_a + B_4x_b + C_4x_c + D_4x_d.$$

We speak of this system as an orthogonal transformation if

$$u_1^2 + u_2^2 + u_3^2 + u_4^2 = x_a^2 + x_b^2 + x_c^2 + x_d^2.$$

If this relation holds good, 4 rules subsume the relations between the linear constants  $A_r, B_r$ , etc. They are as follows :

*Rule of Column Squares.* The sum of the squares of the constants associated with each  $x$ -score is unity, which we write for  $x_n$  when there are 4 independent variables as

$$\sum_{r=1}^{r=4} N_r^2 = 1.$$

*Rule of Column Cross Products.* The sum of the products of the constants associated with any single pair of  $x$ -scores in the same row is zero, which we may write for  $x_n$  and  $x_m$  when there are 4 independent variables as

$$\sum_{r=1}^{r=4} N_r M_r = 0.$$

*Rule of Row Squares.* The sum of the squares of the linear constants definitive of a single  $u$ -score is unity, i.e. for  $u_r$  of a 4-fold set

$$A_r^2 + B_r^2 + C_r^2 + D_r^2 = 1.$$



*Rule of Row Cross Products.* The sum of the products of all pairs of constants definitive of a single pair of  $u$ -scores is zero, i.e. for the 4-fold set

$$A_r A_s + B_r B_s + C_r C_s + D_r D_s = 0.$$

For a system of more than 3 equations the derivation of the rules is very laborious without recourse to determinants; but the student can easily verify them for the 3-fold set if we here give an elementary demonstration for the simplest case, *viz.* :

$$u_1 = A_1 x_a + B_1 x_b; \quad u_2 = A_2 x_a + B_2 x_b \quad \text{and} \quad u_1^2 + u_2^2 = x_a^2 + x_b^2.$$

The rules then take the form :

*Column Squares and Cross Products.*

$$A_1^2 + A_2^2 = 1 = B_1^2 + B_2^2 \quad \text{and} \quad A_1 B_1 + A_2 B_2 = 0.$$

*Row Squares and Row Cross Products.*

$$A_1^2 + B_1^2 = 1 = A_2^2 + B_2^2 \quad \text{and} \quad A_1 A_2 + B_1 B_2 = 0.$$

The derivation is as follows. If  $u_1^2 + u_2^2 = x_a^2 + x_b^2$ ,

$$(A_1^2 + A_2^2)x_a^2 + (B_1^2 + B_2^2)x_b^2 + 2(A_1 B_1 + A_2 B_2)x_a \cdot x_b \equiv x_a^2 + x_b^2.$$

Hence by equating coefficients we derive the *column* rules of (i) above :

$$A_1^2 + A_2^2 = 1 = B_1^2 + B_2^2;$$

$$A_1 B_1 + A_2 B_2 = 0.$$

We now express each  $x$  as the dependent variable by solving the foregoing equations. Thus :

$$A_2 u_1 = A_1 A_2 x_a + A_2 B_1 x_b;$$

$$A_1 u_2 = A_1 A_2 x_a + A_1 B_2 x_b,$$

$$\therefore x_b = \frac{A_2}{A_2 B_1 - A_1 B_2} \cdot u_1 - \frac{A_1}{A_2 B_1 - A_1 B_2} \cdot u_2.$$

Similarly

$$x_a = \frac{B_2}{A_1 B_2 - A_2 B_1} \cdot u_1 - \frac{B_1}{A_1 B_2 - A_2 B_1} \cdot u_2.$$

If  $(A_1 B_2 - A_2 B_1)^{-1} = D$ , we then have

$$x_a = B_2 \cdot D \cdot u_1 - B_1 \cdot D \cdot u_2$$

$$x_b = -A_2 \cdot D \cdot u_1 + A_1 \cdot D \cdot u_2$$

Whence the relation  $(x_a^2 + x_b^2) = (u_1^2 + u_2^2)$  means that

$$(A_2^2 + B_2^2)D^2 \cdot u_1^2 + (A_1^2 + B_1^2)D^2 \cdot u_2^2 - 2(A_1 A_2 + B_1 B_2)D^2 \cdot u_1 u_2 = u_1^2 + u_2^2,$$

$$\therefore A_2^2 + B_2^2 = D^2 = A_1^2 + B_1^2;$$

$$A_1 A_2 + B_1 B_2 = 0.$$

The last equation corresponds to the rule of row cross products. We obtain the rule of row squares as follows :

$$\text{Since } (A_1^2 + A_2^2) = 1 = (B_1^2 + B_2^2),$$

$$2D^2 = (A_1^2 + A_2^2 + B_1^2 + B_2^2) = 2.$$

Whence  $D = 1$  and

$$A_2^2 + B_2^2 = 1 = A_1^2 + B_1^2.$$

The rule of row cross-products has a special statistical meaning, because it defines a necessary condition of the statistical independence of the  $u$ -scores. The reader who is familiar with determinants may generalise the argument which the following illustrates. We suppose that we have a 3-fold sample, and postulate

$$u_1 = A_1 \cdot c_1 + B_1 \cdot c_2 + C_1 \cdot c_3;$$

$$u_2 = A_2 \cdot c_1 + B_2 \cdot c_2 + C_2 \cdot c_3;$$

$$u_3 = A_3 \cdot c_1 + B_3 \cdot c_2 + C_3 \cdot c_3.$$

The cross-product rule for the rows is

$$A_1A_2 + B_1B_2 + C_1C_2 = 0;$$

$$A_1A_3 + B_1B_3 + C_1C_3 = 0;$$

$$A_2A_3 + B_2B_3 + C_2C_3 = 0.$$

If our  $u$ -scores are independent, their covariance is zero. Now both  $u$ -scores and  $c$ -scores are scores with zero mean value, so that

$$\text{Cov}(u_1, u_2) = E(u_1 \cdot u_2) \text{ and } \text{Cov}(c_1, c_2) = E(c_1 \cdot c_2), \text{ etc.,}$$

$$\begin{aligned} \therefore \text{Cov}(u_1, u_2) &= E(A_1c_1 + B_1c_2 + C_1c_3)(A_2c_1 + B_2c_2 + C_2c_3) \\ &= A_1A_2 \cdot E(c_1^2) + B_1B_2 \cdot E(c_2^2) + C_1C_2 \cdot E(c_3^2) + (A_1B_2 + A_2B_1) \cdot E(c_1 \cdot c_2) \\ &\quad + (A_1C_2 + A_2C_1) \cdot E(c_1 \cdot c_3) + (B_1C_2 + B_2C_1) \cdot E(c_2 \cdot c_3). \end{aligned}$$

In these expressions  $E(c_j^2) = 1$ , because the  $c$ -scores are scores of unit variances and  $E(c_p \cdot c_q) = 0$  in virtue of their independence, so that

$$A_1A_2 + B_1B_2 + C_1C_2 = 0. \star \star \star \star$$

*Numerical Illustrations of the u-score transformation.* In 16.01 we defined 3 conditions as sufficient to ensure that the sum of the squares of an  $n$ -fold set of  $u$ -scores each defined in terms of  $n$  unit sample standard scores ( $c_r$ ) by (xii) is a Chi-Square variate of  $n$  degrees of freedom. The general equation for the  $eu$ -score is

$$u_r = A_r c_a + B_r c_b + C_r \cdot c_b \dots \text{etc.}$$

The three conditions are :

- (i) each of the constants  $A_1, B_1, C_1$ , etc. referable to the specification of  $u_1$  has the value  $n^{-\frac{1}{2}}$ ;
- (ii) the sum of the squares of the constants in each row ( $A_r^2, B_r^2, C_r^2$ , etc.) and of the constants in each column (e.g.  $A_1^2, A_2^2, A_3^2 \dots A_n^2$ ) is unity;
- (iii) the total sum of the cross product of corresponding constants in any two columns ( $A_1B_1, A_2B_2$ , etc.) must be zero.

We shall now illustrate the possibility of choosing the constants accordingly by recourse to numerical examples.

*Case I. Two Variables.* Our equations are

$$u_1 = A_1 \cdot c_1 + B_1 \cdot c_2;$$

$$u_2 = A_2 \cdot c_1 + B_2 \cdot c_2.$$

To make the coefficients of  $u_1$  equal to  $n^{-\frac{1}{2}}$ , in which event  $A_1^2 + B_1^2 = 1$ , we put

$$u_1 = \frac{1}{\sqrt{2}}c_1 + \frac{1}{\sqrt{2}}c_2;$$

$$u_2 = A_2 \cdot c_1 + B_2 \cdot c_2.$$



To make  $A_2^2 + B_2^2 = 1$  and  $A_1^2 + A_2^2 = 1 = B_1^2 + B_2^2$ ,  $A_2 = \pm (\frac{1}{2})^{\frac{1}{2}}$  and  $B_2 = \pm (\frac{1}{2})^{\frac{1}{2}}$ ; and the column cross products will vanish if the signs are opposite, when row cross products also vanish. Thus alternative solutions are

$$u_1 = \frac{1}{\sqrt{2}}c_1 + \frac{1}{\sqrt{2}}c_2; \quad u_1 = \frac{1}{\sqrt{2}}c_1 + \frac{1}{\sqrt{2}}c_2;$$

$$u_2 = \frac{1}{\sqrt{2}}c_1 - \frac{1}{\sqrt{2}}c_2; \quad u_2 = -\frac{1}{\sqrt{2}}c_1 + \frac{1}{\sqrt{2}}c_2.$$

*Case II. Three Variables.* We define the constants of the first row to satisfy Rule (i) above, and fix one constant in each remaining row and one other in each column to satisfy Rule (ii). Our equations are then

$$u_1 = \frac{1}{\sqrt{3}}c_1 + \frac{1}{\sqrt{3}}c_2 + \frac{1}{\sqrt{3}}c_3;$$

$$u_2 = A_2 \cdot c_1 + B_2 \cdot c_2 + (1 - A_2^2 - B_2^2)^{\frac{1}{2}}c_3;$$

$$u_3 = \pm (\frac{2}{3} - A_2^2)^{\frac{1}{2}}c_1 \pm (\frac{2}{3} - B_2^2)^{\frac{1}{2}}c_2 \pm (A_2^2 + B_2^2 - \frac{1}{3})^{\frac{1}{2}}c_3.$$

We may satisfy the condition that the cross products of the first two and of the first and the third column vanish by evaluating  $B_2$ , having fixed  $A_2$  arbitrarily. We shall put  $A_2 = 0$ , so that the system of constants becomes

$$\begin{array}{ccc} + (\frac{1}{3})^{\frac{1}{2}} & + (\frac{1}{3})^{\frac{1}{2}} & (\frac{1}{3})^{\frac{1}{2}} \\ 0 & B_2 & \pm (1 - B_2^2)^{\frac{1}{2}} \\ \pm (\frac{2}{3})^{\frac{1}{2}} & \pm (\frac{2}{3} - B_2^2)^{\frac{1}{2}} & \pm (B_2^2 - \frac{1}{3})^{\frac{1}{2}}. \end{array}$$

The cross products of the first and second columns vanish when

$$(\frac{2}{3})^{\frac{1}{2}}(\frac{2}{3} - B_2^2)^{\frac{1}{2}} = -\frac{1}{3}$$

$$\therefore B_2 = \pm (\frac{1}{2})^{\frac{1}{2}}.$$

We have now the set

$$\begin{array}{ccc} + (\frac{1}{3})^{\frac{1}{2}} & + (\frac{1}{3})^{\frac{1}{2}} & + (\frac{1}{3})^{\frac{1}{2}}; \\ 0 & \pm (\frac{1}{2})^{\frac{1}{2}} & \pm (\frac{1}{2})^{\frac{1}{2}}; \\ \pm (\frac{2}{3})^{\frac{1}{2}} & \pm (\frac{1}{6})^{\frac{1}{2}} & \pm (\frac{1}{6})^{\frac{1}{2}}. \end{array}$$

It remains to choose the signs so that all the cross products of the columns vanish, *viz.* :

$$\begin{array}{ccc} + (\frac{1}{3})^{\frac{1}{2}} & + (\frac{1}{3})^{\frac{1}{2}} & + (\frac{1}{3})^{\frac{1}{2}}; \\ 0 & - (\frac{1}{2})^{\frac{1}{2}} & + (\frac{1}{2})^{\frac{1}{2}}; \\ (\frac{2}{3})^{\frac{1}{2}} & - (\frac{1}{6})^{\frac{1}{2}} & - (\frac{1}{6})^{\frac{1}{2}}. \end{array}$$

Our final set of equations is then

$$u_1 = \frac{1}{\sqrt{3}}c_1 + \frac{1}{\sqrt{3}}c_2 + \frac{1}{\sqrt{3}}c_3;$$

$$u_2 = -\frac{1}{\sqrt{2}}c_2 + \frac{1}{\sqrt{2}}c_3;$$

$$u_3 = \frac{\sqrt{2}}{\sqrt{3}}c_1 - \frac{1}{\sqrt{6}}c_2 - \frac{1}{\sqrt{6}}c_3.$$

Again the cross products of any two rows vanish,

$$(A_1A_2 + B_1B_2 + C_1C_2) = 0 = (A_1A_3 + B_1B_3 + C_1C_3) = (A_2A_3 + B_2B_3 + C_2C_3).$$

Another solution which likewise satisfies the prescribed conditions is

$$\begin{array}{ccc} +(\frac{1}{3})^{\frac{1}{2}} & +(\frac{1}{3})^{\frac{1}{2}} & +(\frac{1}{3})^{\frac{1}{2}} \\ +(\frac{1}{2})^{\frac{1}{2}} & -(\frac{1}{2})^{\frac{1}{2}} & 0 \\ +(\frac{1}{6})^{\frac{1}{2}} & +(\frac{1}{6})^{\frac{1}{2}} & -(\frac{2}{3})^{\frac{1}{2}}. \end{array}$$

*Case III. Four Variables.* In this case  $n^{-\frac{1}{2}} = \frac{1}{2}$ , and we may write down our system of constants as

$$\begin{array}{cccc} \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ A_2 & B_2 & C_2 & \pm(1 - A_2^2 - B_2^2 - C_2^2)^{\frac{1}{2}} \\ A_3 & B_3 & C_3 & \pm(1 - A_3^2 - B_3^2 - C_3^2)^{\frac{1}{2}} \\ \pm(\frac{3}{4} - A_2^2 - A_3^2)^{\frac{1}{2}} & \pm(\frac{3}{4} - B_2^2 - B_3^2)^{\frac{1}{2}} & \pm(\frac{3}{4} - C_2^2 - C_3^2)^{\frac{1}{2}} & \pm(A_2^2 + A_3^2 + B_2^2 + B_3^2 + C_2^2 + C_3^2 - \frac{5}{4})^{\frac{1}{2}}. \end{array}$$

We may now fix any 3 constants in the first 2 columns without prejudice to the condition that cross products with other columns must vanish. We shall write  $A_2 = 0 = B_2 = B_3$ ; and our system takes the form

$$\begin{array}{cccc} \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & C_2 & (1 - C_2^2)^{\frac{1}{2}} \\ A_3 & 0 & C_3 & (1 - A_3^2 - C_3^2)^{\frac{1}{2}} \\ \pm(\frac{3}{4} - A_3^2)^{\frac{1}{2}} & \pm(\frac{3}{4})^{\frac{1}{2}} & \pm(\frac{3}{4} - C_2^2 - C_3^2)^{\frac{1}{2}} & (A_3^2 + C_2^2 + C_3^2 - \frac{5}{4})^{\frac{1}{2}}. \end{array}$$

To satisfy the condition that cross products of the first two columns vanish

$$\begin{aligned} \frac{3}{4}(\frac{3}{4} - A_3^2) &= (-\frac{1}{4})^2 \\ \therefore A_3 &= \pm(\frac{2}{3})^{\frac{1}{2}}. \end{aligned}$$

The sign of  $A_3$  is immaterial but  $(\frac{3}{4} - A_3^2)^{\frac{1}{2}}$  and  $(\frac{3}{4})^{\frac{1}{2}}$  in the bottom row must have opposite signs, and we may put

$$\begin{array}{cccc} \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & C_2 & (1 - C_2^2)^{\frac{1}{2}} \\ +(\frac{2}{3})^{\frac{1}{2}} & 0 & C_3 & (\frac{1}{3} - C_3^2)^{\frac{1}{2}} \\ +(\frac{1}{12})^{\frac{1}{2}} & -(\frac{3}{4})^{\frac{1}{2}} & \pm(\frac{3}{4} - C_2^2 - C_3^2)^{\frac{1}{2}} & (C_2^2 + C_3^2 - \frac{7}{12})^{\frac{1}{2}}. \end{array}$$

We now fix  $C_2$  and  $C_3$  so that the cross products of the first and third and of the second and third columns vanish. By taking the third with the second we get

$$\begin{aligned} \frac{3}{4}(\frac{3}{4} - C_2^2 - C_3^2) &= \frac{1}{16} \\ \therefore (\frac{3}{4} - C_2^2 - C_3^2)^{\frac{1}{2}} &= \pm(\frac{1}{12})^{\frac{1}{2}}. \end{aligned}$$

Without prejudice we may fix the sign of the above, and obtain from the first and third columns

$$\begin{aligned} \frac{1}{4} + (\frac{2}{3})^{\frac{1}{2}}C_3 + \frac{1}{12} &= 0 \\ \therefore C_3 &= -(\frac{1}{6})^{\frac{1}{2}}. \end{aligned}$$

Whence from the foregoing expression for  $(\frac{3}{4} - C_2^2 - C_3^2)^{\frac{1}{2}}$

$$\frac{7}{12} - C_2^2 = \frac{1}{12} \quad \text{and} \quad C_2 = \pm(\frac{1}{2})^{\frac{1}{2}}.$$









It will suffice to consider the statistics  $s_c^2$  and  $s_r^2$  for a set-up involving 2 criteria of classification as in 13.03 and 13.04. If  $M_i$  is a column mean and  $M_x$  is the grand mean of the sample, we define  $s_c^2$  by the relation

$$s_c^2 = \frac{r \sum_{i=1}^{i=c} (M_i - M_x)^2}{c - 1} \quad . \quad . \quad . \quad . \quad . \quad . \quad (i)$$

If  $\sigma^2$  is the true variance of the score distribution, that of the mean of the  $r$ -fold column sample is given by

$$\sigma_c^2 = \frac{\sigma^2}{r} \quad . \quad . \quad . \quad . \quad . \quad . \quad (ii)$$

Whence we have

$$\frac{s_c^2}{\sigma^2} = \frac{\sum_{i=1}^{i=c} (M_i - M_x)^2}{(c - 1)\sigma_c^2} \quad . \quad . \quad . \quad . \quad . \quad . \quad (iii)$$

We shall write  $S_i = (c - 1)s_c^2$ , so that

$$\frac{S_i}{\sigma^2} = \sum_{i=1}^{i=c} \frac{(M_i - M_x)^2}{\sigma_c^2} \quad . \quad . \quad . \quad . \quad . \quad . \quad (iv)$$

If  $M$  is the true mean score, we may write as in 16.01

$$\frac{S_i}{\sigma^2} = \sum_{i=1}^{i=c} \frac{(M_i - M)^2}{\sigma_c^2} - \frac{c(M_x - M)^2}{\sigma_c^2} \quad . \quad . \quad . \quad . \quad . \quad . \quad (v)$$

In this expression, we have

$$M_x = \sum_{i=1}^{i=c} \frac{M_i}{c} \quad . \quad . \quad . \quad . \quad . \quad . \quad (vi)$$

Thus  $M_x$  is the mean of a  $c$ -fold sample of column mean scores. Hence the variance of the distribution of  $M_x$  is given by

$$\sigma_x^2 = \frac{\sigma_c^2}{c},$$

$$\therefore \frac{S_i}{\sigma^2} = \sum_{i=1}^{i=c} \frac{(M_i - M)^2}{\sigma_c^2} - \frac{(M_x - M)^2}{\sigma_x^2} \quad . \quad . \quad . \quad . \quad . \quad . \quad (vii)$$

If the score distribution is normal, that of the mean is also normal. Hence we have now expressed (iv) as the difference between a  $c$ -fold sum of square normal scores of unit variance and one square normal score of unit variance, i.e. as the difference between a Chi-Square variate of  $c$  degrees of freedom and a Chi-Square variate of 1 degree of freedom like the expression defined by (vi) of 16.01. In (vii) above  $M_x$  is the mean of  $M_i$  as  $M_x$  is the mean of  $M_j$  in (viii). Thus the two expressions are in all respects comparable. It is therefore unnecessary to repeat with appropriate change of symbols the orthogonal transformation of 16.01 in order to show that  $(S_i \div \sigma^2)$  is a Chi-Square variate of  $c - 1$  degrees of freedom.

*Mutatis mutandis* the same argument applies to the statistic

$$(r - 1)s_r^2 = S_j = c \sum_{j=1}^{j=r} (M_j - M_x)^2 \quad . \quad . \quad . \quad . \quad . \quad . \quad (viii)$$

In this case  $(S_j \div \sigma^2)$  is a Chi-Square variate of  $(r - 1)$  degrees of freedom.

The total sample variance yields an unbiased estimate of the same form as (v) and (vi) in 16.01 based on  $rc$ -scores. We may write it as

$$(rc - 1)s_i^2 = S_{ij} = \sum_{i=1}^{i=c} \sum_{j=1}^{j=r} (x_{ij} - M_x)^2 \quad . \quad . \quad . \quad . \quad (ix)$$

As in the derivation of (v) we obtain

$$\frac{S_{ij}}{\sigma^2} = \sum_{i=1}^{i=c} \sum_{j=1}^{j=r} \frac{(x_{ij} - M)^2}{\sigma^2} - \frac{rc(M_x - M)^2}{\sigma^2} \quad . \quad . \quad . \quad . \quad (x)$$

Since  $M_x$  is the mean of an  $rc$ -fold sample, the variance of its distribution is given by

$$\sigma_x^2 = \frac{\sigma^2}{rc},$$

$$\therefore \frac{S_{ij}}{\sigma^2} = \sum_{i=1}^{i=c} \sum_{j=1}^{j=r} \frac{(x_{ij} - M)^2}{\sigma^2} - \frac{(M_x - M)^2}{\sigma_x^2} \quad . \quad . \quad . \quad . \quad (xi)$$

It is thus evident that  $S_{ij}$  in standard form is a Chi-Square variate of  $(rc - 1)$  d.f.

That the numerator of the estimate  $s_z^2$  of 13.03 and 13.04 is also in standard form a Chi-Square variate with  $(r - 1)(c - 1)$  degrees of freedom is less easy to see, and calls for closer examination. We define it by the relations

$$s_z^2 = \frac{rc \cdot V_z}{(r - 1)(c - 1)};$$

$$M(V_r) + M(V_c) - V_x = V_z = V_x - V(M_r) - V(M_c).$$

We may write the preceding expression more fully in standard form as

$$\frac{rc \cdot V_z}{\sigma^2} = \sum_{i=1}^{i=c} \sum_{j=1}^{j=r} \frac{(x_{ij} - M_x)^2}{\sigma^2} - \sum_{j=1}^{j=r} c \cdot \frac{(M_j - M_x)^2}{\sigma^2} - \sum_{i=1}^{i=c} r \cdot \frac{(M_i - M_x)^2}{\sigma^2}.$$

We may transform this as above to 4 sets of square normal scores of unit variance, viz. :

$$\begin{aligned} \frac{(r - 1)(c - 1)}{\sigma^2} s_z^2 = \frac{rc \cdot V_z}{\sigma^2} &= \sum_{i=1}^{i=c} \sum_{j=1}^{j=r} \frac{(x_{ij} - M)^2}{\sigma^2} + \frac{(M_x - M)^2}{\sigma_x^2} \\ &\quad - \sum_{j=1}^{j=r} \frac{(M_j - M)^2}{\sigma_r^2} - \sum_{i=1}^{i=c} \frac{(M_i - M)^2}{\sigma_c^2} \quad . \quad (xii) \end{aligned}$$

It will be more easy to follow the appropriate orthogonal transformation, if we drop the subscript symbolism, and deal with the procedure. Accordingly, we assume the following schema of independent normal score values :

Scores	Scores				Mean
	$x_1$	$x_2$	$x_3$	$x_4$	$M_{1 \cdot r}$
	$x_5$	$x_6$	$x_7$	$x_8$	$M_{2 \cdot r}$
	$x_9$	$x_{10}$	$x_{11}$	$x_{12}$	$M_{3 \cdot r}$
Mean	$M_{1 \cdot c}$	$M_{2 \cdot c}$	$M_{3 \cdot c}$	$M_{4 \cdot c}$	$M_x$



For this assemblage we may write the first term of the expression on the right of (xii) in the form

$$\sum_{i=1}^c \sum_{j=1}^r \frac{(x_{ij} - M)^2}{\sigma^2} = \sum_{s=1}^{s=12} \left[ \frac{x_s - M}{\sigma} \right]^2 = \sum_{s=1}^{s=12} c_s^2 \quad . \quad . \quad . \quad (xiii)$$

By definition

$$12M_x = \sum_{s=1}^{s=12} x_s \quad \text{and} \quad 12(M_x - M) = \sum_{s=1}^{s=12} (x_s - M),$$

$$\therefore \frac{12(M_x - M)}{\sigma} = \sum_{s=1}^{s=12} c_s.$$

Since  $\sigma = \sqrt{12}\sigma_x$

$$\frac{M_x - M}{\sigma_x} = \frac{1}{\sqrt{12}} \sum_{s=1}^{s=12} c_s \quad . \quad . \quad . \quad . \quad (xiv)$$

We shall now assume that it is possible to transform the  $c$ -scores into  $u$ -scores in accordance with the orthogonal relation

$$\sum_{s=1}^{s=12} u_s^2 = \sum_{s=1}^{s=12} c_s^2 \quad . \quad . \quad . \quad . \quad . \quad (xv)$$

In virtue of (xiv), we are then free to define

$$u_1 = \frac{M_x - M}{\sigma_x} = \frac{1}{\sqrt{12}} \sum_{s=1}^{s=12} c_s \quad . \quad . \quad . \quad . \quad . \quad (xvi)$$

Hence we have the following expression for the first two terms on the right hand side of (xii):

$$\sum_{j=1}^r \sum_{i=1}^c \frac{(x_{ij} - M)^2}{\sigma^2} + \frac{(M_x - M)^2}{\sigma_x^2} = 2u_1^2 + \sum_{s=2}^{s=12} u_s^2 \quad . \quad . \quad . \quad (xvii)$$

We shall now write in the term of (xii) involving the row means

$$\frac{(M_{1..} - M)}{\sigma_r} = v_1; \quad \frac{(M_{2..} - M)}{\sigma_r} = v_2; \quad \frac{(M_{3..} - M)}{\sigma_r} = v_3 \quad . \quad . \quad (xviii)$$

Similarly, for the term involving the column means we shall put

$$\begin{aligned} \frac{(M_{1.c} - M)}{\sigma_c} &= w_1; \quad \frac{(M_{2.c} - M)}{\sigma_c} = w_2; \\ \frac{(M_{3.c} - M)}{\sigma_c} &= w_3; \quad \frac{(M_{4.c} - M)}{\sigma_c} = w_4 \quad . \quad . \quad . \quad . \quad (xix) \end{aligned}$$





This expression involves  $6 = 2.3 = (r-1)(c-1)$  square  $u$ -scores, i.e.  $(r-1)(c-1)$  independent Chi-Square variates of 1 d.f. The general pattern of the transformation is as follows:

$$\begin{aligned} \frac{(r-1)(c-1)}{\sigma^2} s_z^2 &= \sum_{s=1}^{s=rc} u_s^2 + u_1^2 - \left[ 2u_1^2 + \sum_{s=2}^{s=r} u_s^2 + \sum_{s=r+1}^{s=r+c-1} u_s^2 \right] \\ &= \sum_{s=2}^{s=rc} u_s^2 - \sum_{s=2}^{s=r+c-1} u_s^2 \\ &= \sum_{s=r+c}^{s=rc} u_s^2. \end{aligned}$$

The last expression contains  $rc - (r+c) + 1 = (r-1)(c-1)$  terms.

At this point, the reader may reasonably want assurance that the simultaneous transformation of  $(r+c-1)$  of the  $u$ -scores into  $v$ -scores and  $w$ -scores in the foregoing is consistent with the orthogonal relation between the entire  $rc$ -fold set-up of  $u$ -scores and  $c$ -scores. In accordance with results obtained in 16.02, the following arbitrary constants satisfy the  $v$  and  $w$  score transformations:

$$\begin{aligned} u_1 &= \frac{1}{\sqrt{3}}v_1 + \frac{1}{\sqrt{3}}v_2 + \frac{1}{\sqrt{3}}v_3; \\ u_2 &= \frac{1}{\sqrt{2}}v_1 - \frac{1}{\sqrt{2}}v_2 \dots; \\ u_3 &= \frac{1}{\sqrt{6}}v_1 + \frac{1}{\sqrt{6}}v_2 - \frac{1}{\sqrt{3}}v_3; \\ * \quad * \quad * \quad * \quad * \quad * \quad * \quad * \\ u_1 &= \frac{1}{2}w_1 + \frac{1}{2}w_2 + \frac{1}{2}w_3 + \frac{1}{2}w_4; \\ u_4 &= \frac{1}{\sqrt{2}}w_1 - \frac{1}{\sqrt{2}}w_2 \dots; \\ u_5 &= \dots \frac{1}{\sqrt{2}}w_3 - \frac{1}{\sqrt{2}}w_4; \\ u_6 &= \frac{1}{2}w_1 + \frac{1}{2}w_2 - \frac{1}{2}w_3 - \frac{1}{2}w_4. \end{aligned}$$

It will suffice to examine the orthogonal property of the pair  $u_3$  and  $u_6$  so defined when we express them in terms of the  $c$ -scores. From (xxii) and (xxv), we have

$$\begin{aligned} u_3 &= \frac{1}{\sqrt{24}}(c_1 + c_2 + c_3 + c_4 + c_5 + c_6 + c_7 + c_8) - \frac{1}{\sqrt{6}}(c_9 + c_{10} + c_{11} + c_{12}); \\ u_6 &= \frac{1}{\sqrt{12}}(c_1 + c_2 + c_5 + c_6 + c_9 + c_{10}) - \frac{1}{\sqrt{12}}(c_3 + c_4 + c_7 + c_8 + c_{11} + c_{12}). \end{aligned}$$

Evidently the sum of the squares of the constants in each row is unity and the sum of the cross products vanishes as we see if we set them out as below:

	$c_1$	$c_2$	$c_3$	$c_4$	$c_5$	$c_6$	$c_7$	$c_8$	$c_9$	$c_{10}$	$c_{11}$	$c_{12}$
$u_3$	$\frac{1}{\sqrt{24}}$	$\frac{1}{\sqrt{24}}$	$\frac{1}{\sqrt{24}}$	$\frac{1}{\sqrt{24}}$	$\frac{1}{\sqrt{24}}$	$\frac{1}{\sqrt{24}}$	$\frac{1}{\sqrt{24}}$	$\frac{1}{\sqrt{24}}$	$-\frac{1}{\sqrt{6}}$	$-\frac{1}{\sqrt{6}}$	$-\frac{1}{\sqrt{6}}$	$-\frac{1}{\sqrt{6}}$
$u_6$	$\frac{1}{\sqrt{12}}$	$\frac{1}{\sqrt{12}}$	$-\frac{1}{\sqrt{12}}$	$-\frac{1}{\sqrt{12}}$	$\frac{1}{\sqrt{12}}$	$\frac{1}{\sqrt{12}}$	$-\frac{1}{\sqrt{12}}$	$-\frac{1}{\sqrt{12}}$	$\frac{1}{\sqrt{12}}$	$\frac{1}{\sqrt{12}}$	$-\frac{1}{\sqrt{12}}$	$-\frac{1}{\sqrt{12}}$

For this set-up we can thus define 6  $u$ -scores as linear functions of the  $c$ -scores in conformity with the rules: (a) that the sum of the cross products vanishes for any pair of rows; (b) that the sum of the squares of the constants in any one row is unity. It remains to define 6  $u$ -scores

$$\begin{aligned} & \frac{\sum_{i=1}^{i=c} \sum_{j=1}^{j=r} (x_{ij} - M)^2}{\sigma_x^2} + \frac{(M_x - M)^2}{\sigma_x^2} - \frac{\sum_{j=1}^{j=r} (M_j - M)^2}{\sigma_r^2} - \frac{\sum_{i=1}^{i=c} (M_i - M)^2}{\sigma_c^2} \\ &= \sum_{m=1}^{m=rc} u_m^2 + u_1^2 - \sum_{m=1}^{m=r} u_m^2 - u_1^2 - \sum_{m=r+1}^{m=r+c-1} u_m^2 \\ &= \sum_{m=1}^{m=rc} u_m^2 - \sum_{m=1}^{m=r+c-1} u_m^2 \\ &= \sum_{m=r+c}^{m=rc} u_m^2, \text{ with } (rc - r - c + 1) = (r-1)(c-1) \text{ terms.} \end{aligned}$$



When there are two columns  $(c - 1) = 1$ ; and we shall now see that the ratio of  $s_c^2$  to the true variance of the score distribution is a Chi-Square variate for 1 d.f. If we put  $r_1 = a$ ,  $r_2 = b$  and  $(a + b) = n$ , we may then write

$$\frac{s_c^2}{\sigma^2} = \frac{a(M_a - M_x)^2}{\sigma^2} + \frac{b(M_b - M_x)^2}{\sigma^2} \quad . \quad . \quad . \quad . \quad (\text{xxvii})$$

In this expression the sample mean is given by

$$M_x = \frac{aM_a + bM_b}{n} . . . . . (\text{xxviii})$$

Whence we have

$$\begin{aligned} a(M_a - M_x)^2 + b(M_b - M_x)^2 &= a \cdot M_a^2 + b \cdot M_b^2 + nM_x^2 - 2M_x(a \cdot M_a + b \cdot M_b) \\ &= a \cdot M_a^2 + b \cdot M_b^2 + nM_x^2 - 2nM_x^2, \\ \therefore a(M_a - M_x)^2 + b(M_b - M_x)^2 &= a \cdot M_a^2 + b \cdot M_b^2 - n \cdot M_x^2 \quad . \quad . \quad . \quad (\text{xxix}) \end{aligned}$$

If the unknown true mean score is  $M$ , we may put

$$\begin{aligned} a(M_a - M)^2 + b(M_b - M)^2 &= a \cdot M_a^2 + b \cdot M_b^2 - 2M(a \cdot M_a + b \cdot M_b) + nM^2 \\ &= a \cdot M_a^2 + b \cdot M_b^2 - 2nM \cdot M_x + nM^2 \\ &= (a \cdot M_a^2 + b \cdot M_b^2 - nM_x^2) + (nM_x^2 - 2nM \cdot M_x + nM^2). \end{aligned}$$

Whence from (xxix)

$$\begin{aligned} a(M_a - M)^2 + b(M_b - M)^2 &= a(M_a - M_x)^2 + b(M_b - M_x)^2 + n(M_x - M)^2, \\ \therefore a(M_a - M_x)^2 + b(M_b - M_x)^2 &= a(M_a - M)^2 + b(M_b - M)^2 - n(M_x - M)^2, \\ \therefore \frac{s_c^2}{\sigma^2} &= \frac{a(M_a - M)^2}{\sigma^2} + \frac{b(M_b - M)^2}{\sigma^2} - \frac{n(M_x - M)^2}{\sigma^2} \quad \cdot \quad \cdot \quad \cdot \quad (\text{xxx}) \end{aligned}$$

If we denote the variances of the distributions of the column means by  $\sigma_a^2$  and  $\sigma_b^2$  and that of the grand mean by  $\sigma_x^2$ , we have

$$a \cdot \sigma_a^2 = \sigma^2 = b \cdot \sigma_b^2 = n\sigma_x^2 \quad . \quad . \quad . \quad . \quad . \quad (\text{xxxi})$$

Each of the three terms on the right of (xxx) is thus a square score of unit variance, viz.:

$$\frac{s_c^2}{\sigma^2} = \frac{(M_a - M)^2}{\sigma_a^2} + \frac{(M_b - M)^2}{\sigma_b^2} - \frac{(M_x - M)^2}{\sigma_x^2} \quad \cdot \quad \cdot \quad \cdot \quad (\text{xxxii})$$

We shall now assume that the score distribution is normal whence those of the column sample means and that of the grand mean are normal, so we may write in the usual way as square standard normal scores

$$c_a^2 = \frac{(M_a - M)^2}{\sigma_a^2} \quad \text{and} \quad c_b^2 = \frac{(M_b - M)^2}{\sigma_b^2} \quad . \quad . \quad . \quad (\text{xxxiii})$$

In the third term of (xxxii), we note that

$$\therefore \frac{(M_x - M)}{\sigma_x} = \frac{a(M_a - M)}{n\sigma_x} + \frac{b(M_b - M)}{n\sigma_x}.$$





When our concern is with only one criterion of classification, the numbers of items ( $r_i$ ) in each column need not be the same, and if  $n$  is the total number of cell entries, i.e.  $n = rc$  in the above,

$$\frac{n-c}{n}s_d^2 = M(V_c) \quad \text{and} \quad \frac{(n-c)s_d^2}{\sigma^2} = \sum_{i=1}^{i=c} \sum_{j=1}^{j=r_i} \frac{(x_{ij} - M_i)^2}{\sigma^2}. \quad (\text{xxxviii})$$

## 16.05 THE PAIRED DIFFERENCE TEST FOR SMALL SAMPLES

In Chapter 7 of Vol. I we have distinguished between two ways of investigating a real difference in the domain of representative scoring :

- (a) comparison of the mean scores of groups subjected to different treatments as in 16.06 below ;
- (b) comparison of response of one and the same individual before and after treatment or of the effect of different treatments on pairs of individuals the two members of which share a common peculiarity.

The second procedure involves the null hypothesis that the true mean difference between paired scores is zero ; and we can test its validity if entitled to assume that the  $d$ -score (i.e. paired score difference) distribution is normal, as it will be if we regard each pair of observations as a 2-fold sample from a normal universe. On that assumption, we can safely apply the  $c$ -test for a normally distributed score, if the sample is large. If  $M_d$  is the mean difference and  $s_m^2$  is an unbiased estimate of the true variance  $\sigma_m^2$  of its distribution, the appropriate ratio for a  $p$ -fold sample of paired scores is as given by (xix) in 14.08 :

[illegible]

In view of what follows, it is important to stress that  $s_m^2$  is an unbiased estimate of  $\sigma_m^2$  as defined by the relations implicit in the above, *viz.* :

$$E_s(s_m^2) = \sigma_m^2 \quad \text{and} \quad s_m^2 = \frac{1}{p^2} \sum_{r=1}^{r=p} d_r^2 . \qquad . \qquad . \qquad . \qquad (\text{ii})$$

If  $M$  is the true mean and  $\sigma_d^2$  is the unknown true variance of the  $d$ -score distribution, we may write in the symbolism of 13.02:

$$E_s \cdot E_r (d_r - M)^2 = \sigma_d^2 = E_s \cdot E_r (d_r^2) - M^2.$$

In this case, the null hypothesis implies that  $M = 0$ ,

$$\therefore \frac{1}{p} E_s \cdot E_r(d_r^2) = \frac{\sigma_d^2}{p} = \sigma_m^2.$$

In the preceding expression

$$E_r(d_r^2) = \frac{1}{p} \sum_{r=1}^{r=p} d_r^2.$$

Whence as in (ii) :

$$E_s(s_m^2) = \sigma_m^2.$$

$$\frac{t^2}{r} = \frac{u_1^2}{u_2^2 + u_3^2 + \dots + u_n^2}.$$



In the above our  $u$ -scores are by definition independent, and we have eliminated  $u_1^2$  from the denominator. Hence the numerator and denominator are independent statistics, i.e.  $(t^2 \div f)$  is the ratio of a Chi-Square variate of 1 d.f. to an independent Chi-Square variate of  $f$  degrees of freedom.

Now we have seen in 15.04 that if  $x$  is the ratio of a Chi-Square variate of 1 d.f. to an independent Chi-Square variate of  $f$  degrees of freedom, its p.d. is

$$f(x) = \frac{1}{B(\frac{1}{2}, \frac{1}{2}f)} \cdot \frac{x^{-\frac{1}{2}}}{(1+x)^{\frac{1}{2}(f+1)}}.$$

In this case  $x = (t^2 \div f)$  and the simple scalar substitution of Case I in 15.03 yields

$$F(t^2) = \frac{1}{B(\frac{1}{2}, \frac{1}{2}f)\sqrt{f}} \cdot \frac{t^{-1}}{\left(1 + \frac{t^2}{f}\right)^{\frac{1}{2}(f+1)}} \quad . \quad . \quad . \quad . \quad (\text{vi})$$

The derivation of the p.d. of  $t$  itself in accordance with Case IV of 15.03 presupposes an ulterior reason for believing that  $f(t)$  is a symmetrical function of  $t$ . The latter is the ratio of the deviation of the sample mean from the true mean to the square root of the sample estimate of its variance; and the distribution of this ratio is necessarily symmetrical if the distribution of score deviations in the parent universe is itself symmetrical. This is easy to see of the discrete universe and hence of a hypothetical continuous universe in the limit. For a given numerical positive value of the score-sum or mean ( $M$ ) we may pair off every uniquely constituted combination of unit scores with an otherwise identical set of reverse sign, their mean ( $-M$ ) being numerically equivalent to  $M$  but negative.

One illustration which the reader may explore more fully should suffice to make this clear, *viz.* extraction of 3-fold samples from a 7-class rectangular universe of score deviations  $-3, -2, -1, 0, +1, +2, +3$ . We need **only** consider score-sums of  $\pm 3$  with mean value  $\pm 1$ . All corresponding combinations of unit scores consistent with these values are then

$$\begin{array}{ll} (M = +1) & +3, 0, 0 \dots +3, -1, +1 \dots +3, -2, +2 \dots \\ & \phantom{+3, 0, 0 \dots} +2, +1, 0 \dots +2, +2, -1 \dots +1, +1, +1, \\ (M = -1) & -3, 0, 0 \dots -3, +1, -1 \dots -3, +2, -2 \dots \\ & \phantom{-3, 0, 0 \dots} -2, -1, 0 \dots -2, -2, +1 \dots -1, -1, -1. \end{array}$$

Thus corresponding sample values of  $+M$  and  $-M$  will have equal frequency if corresponding negative and positive values of the unit scores ( $x_r$ ) have equal frequency; and sample variances corresponding to each pair of numerically identical scores of reverse sign will be necessarily identical, and will not affect the ratio of  $(M_x - M)$  to  $s_m$ . The distribution of the ratio therefore depends only on the distribution of  $M_x$ ; and this is necessarily symmetrical, if the distribution of  $x$ -scores in the parent universe is symmetrical, as is true of the normal universe we postulate in this context.

As in Example 5 of 15.03, we may thus write

$$g(t) = \frac{1}{B(\frac{1}{2}, \frac{1}{2}f)\sqrt{f}} \cdot \frac{1}{\left(1 + \frac{t^2}{f}\right)^{\frac{1}{2}(f+1)}} \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \text{(vii)}$$

The table of the  $t$ -integral in Kendall's treatise gives the probability ( $P_t$ ) of a value being numerically as great as or greater than  $\pm t$ , i.e.

$$1 - P_t = \int_{-t}^t g(t) \cdot dt.$$





$$\begin{aligned} t^2 &= \frac{n(n-1)M_d^2}{S_n - n.M_d^2} \quad \text{and} \quad c^2 = \frac{n^2 M_d^2}{S_n}, \\ \therefore \frac{c^2}{t^2} &= \frac{n}{n-1} \left\{ 1 - \frac{n.M_d^2}{S_n} \right\} = \frac{(f+1) - c^2}{f}, \\ \therefore c^2 &= \frac{(f+1)t^2}{f+t^2}. \end{aligned}$$

$f$	$t = 2$			$t = 3$		
	$c$	$P_c$	$P_t$	$c$	$P_c$	$P_t$
5	1.633	0.103	0.102	1.964	0.0496	0.0300
10	1.773	0.076	0.074	2.283	0.0224	0.0134
15	1.835	0.067	0.064	2.449	0.0143	0.0090
20	1.871	0.061	0.060	2.553	0.0107	0.0071
30	1.910	0.056	0.055	2.675	0.0075	0.0054
40	1.931	0.054	0.052	2.744	0.0061	0.0046
60	1.953	0.051	0.050	2.821	0.0048	0.0039

$$\begin{aligned} \sum d^2 &= 296 ; M_d = 1.33 ; \sum d = 52 ; \\ \sum (d - M_d)^2 &= \sum d^2 - n . M_d^2 = 296 - 69.3 = 226.7, \\ \therefore t^2 &= \frac{38.39.(1.33)^2}{226.7} = \frac{2633.4}{226.7}, \\ \therefore t &\simeq 3.4. \end{aligned}$$
$$c \simeq 3.1.$$

\* \* \* \* \*

*Confidence Limits of an Estimated Mean.* There is, however, another and important use for the  $t$ -distribution interpreted as that of the deviation of the mean ( $M_x$ ) of the  $n$ -fold sample from the true mean ( $M$ ) of a homogeneous normal parent universe. It may happen that we want to estimate  $M$  in which case  $M_x$  is an unbiased statistic ; but it is possible to take a step further, i.e. to assign confidence limits between which  $M$  lies.

We shall see how to do this more readily, if we first assume that we know the value of  $M$  and  $\sigma^2$ , the variance of the u.s.d., whence also  $\sigma_m^2 = (\sigma^2 \div n)$ . We may denote the deviation of the sample mean from the true mean by  $h = (M_x - M)$ . By hypothesis, therefore  $(h \div \sigma_m) = c$  is a normal score of unit variance. At the  $2\sigma$  level  $h = 2\sigma_m$  and  $c = 2$ . The probability that  $h$  will be numerically equal to or greater than  $2\sigma_m$  is then given by the table of the normal integral as

$$P_2 = \sqrt{\frac{2}{\pi}} \int_0^2 e^{-\frac{1}{2}c^2} dc \simeq 0.945.$$

Let us now suppose that we know the value of  $\sigma^2$  and hence of  $\sigma_m^2$  but that we do not know the value of  $M$ . If  $(M_x - M)$  lies within the limits  $\pm h = \pm 2\sigma_m$ , it follows that  $M$  lies within the limits  $(M_x \mp h) = (M_x \mp 2\sigma_m)$ . Either statement is true of 95 per cent. of all samples we meet, and we shall therefore err in only 5 per cent. of our samples, if we consistently set our estimate of  $M$  in the range from  $M_x - 2\sigma_m$  to  $M_x + 2\sigma_m$ . For instance, we may suppose the sample mean is 11.5, and that the true value of  $\sigma_m$  is 0.75, so that  $2\sigma_m = 1.5$ . A deviation of 1.5 either way signifies in this case that  $\pm (11.5 - M) = \pm 1.5$ , whence that  $M$  lies within the range from 10 to 13 inclusive. Such then are the limits of the range of admissible values of  $M$  at the 95 per cent. confidence level.

The foregoing argument presupposes that we know the value of  $\sigma_m$ . This will rarely if ever be true in laboratory or field work, though it is easy to construct a model set-up in which it would be so. A more usual type of situation is that of the investigator who wishes to assign a value to the length of a piece of wire on the basis of successive observations ( $x$ ) subject to an approximately normal distribution of instrumental error. He then has two sample parameters on which to base his judgment, *viz.* the sample mean  $M_x$  and an *estimate* of the variance ( $\sigma_m$ ) of the mean, *viz.* :

$$s_m^2 = \sum_{r=1}^{r=n} \frac{(x_r - M_x)^2}{n(n-1)}.$$

The  $t$ -distribution then defines that of the ratio  $(M_x - M) \div s_m$ . For a particular value of  $f = (n - 1)$  the table of the  $t$ -integral cites how large  $t$  must be if  $P \simeq 0.05$  is the probability that the value of  $t$  lies outside a prescribed numerical value. Let us suppose that the table cites  $t = \pm a$  as the prescribed value, so that

$$M_x - M = \pm a s_m \text{ and } M = M_x \pm a s_m.$$

We can then say that  $M$  lies within the limits  $M_x \pm a \cdot s_m$  at the 95 per cent. confidence level, if we assign to  $a$  the tabular value prescribed by  $P \simeq 0.05$ .

For simplicity, we may take the foregoing figure for the observed mean  $M_x = 11.5$  and assume that  $s_m = 0.75$  for a sample of 10, so that  $f = 9$ . For this value of  $f$  the table of the  $t$ -integral gives  $t = 2.26$  at the level  $P = 0.05$ , i.e. odds of about 20 : 1 against getting a value of  $t$  numerically equal to or greater than 2.26. This specifies a deviation  $\pm (2.26)(0.75) = \pm 1.695$ . At the 95 per cent. confidence level we shall thus say that the true mean will lie in the range  $11.5 \pm 1.695$  or from 9.805 to 13.195. For  $f > 50$  the normal integral will give a figure which does not appreciably differ from the result of proceeding as in this example. The reader should be able to interpret the appropriate procedure for any other confidence level (e.g. 99 per cent.).

#### 16.06 THE GROUP MEAN DIFFERENCE TEST

In contradistinction to the approximate  $c$ -test of 7.06 in Vol. I we have examined in 13.07 an alternative approach to the recognition of a difference between the mean score of two



independent samples. If one sample consists of  $a$  and the other of  $b = (n - a)$  items, we may define two statistics as in 16.04 by the relations

[illegible]

$$s_a^2 = \frac{1}{n-2} \sum_{j=1}^{j=a} (x_{aj} - M_a)^2 + \frac{1}{n-2} \sum_{j=1}^{j=b} (x_{bj} - M_b)^2 = s_a^2 + s_b^2 \quad . \quad . \quad (ii)$$

In the last expression

$$s_a^2 = \frac{1}{n-2} \sum_{j=1}^{j=a} (x_{aj} - M_a)^2 \quad \text{and} \quad s_b^2 = \frac{1}{n-2} \sum_{j=1}^{j=b} (x_{bj} - M_b)^2.$$

As we have seen in 13.07 the ratio of the two is then equivalent to

$$\frac{s_c^2}{s_d^2} = \frac{(M_a - M_b)^2}{\left(\frac{1}{a} + \frac{1}{b}\right)(s_a^2 + s_b^2)} = \frac{ab(M_a - M_b)^2}{n(s_a^2 + s_b^2)} \quad \text{. . . . . (iii)}$$

The statistics denoted by  $s_c^2$  and  $s_d^2$  are each estimates of the true variance ( $\sigma^2$ ) of the score distribution of the putative common universe of the null hypothesis, i.e. that the column samples do in fact come from one and the same universe. Their consistency is therefore a criterion of the absence of a difference between the column means other than such as might arise by random sampling. We may express the ratio defined by (iii) as

$$\therefore R^2 = \frac{s_c^2}{\sigma_d^2} \div \frac{(n-2)s_d^2}{\sigma_d^2} \quad \text{. . . . . (iv)}$$

From (xxxii) and (xxxvi) of 16.04 we can see that  $R^2$  expressed in this form is the ratio of a Chi-Square variate of 1 d.f. to a Chi-Square variate of  $f = (n - 2)$  degrees of freedom. The problem of the distribution of  $R$  is therefore soluble, if we can show that these are statistically independent. We first recall a result obtained in 16.05, where we have seen that it is possible to express the numerator of (iv) in terms of two independent normal scores of unit variance, *viz.* :

$$\frac{s_c^2}{\sigma^2} = v_1^2 + v_2^2 - \left[ \left( \frac{a}{n} \right)^{\frac{1}{2}} v_1 + \left( \frac{b}{n} \right)^{\frac{1}{2}} v_2 \right]^2 \quad (v)$$

In (v) the meaning of  $v_1$  and  $v_2$  is

$$v_1 = \frac{(M_a - M)}{\sigma_a} \quad \text{and} \quad v_2 = \frac{(M_b - M)}{\sigma_b}.$$

We may transform the denominator in (iv) as follows :

[illegible]





From the above we have  $a = 25 = b$  and  $n - 2 = 48$ . We denote the boy's score by  $x_a$  and the girl's by  $x_b$ :

$$\begin{aligned} \sum x_a &= 197.50; & M_a &= 7.90; & aM_a^2 &= 1560.25; \\ \sum x_a^2 &= 1569.25; & \sum (x_a - M_a)^2 &= \sum x_a^2 - aM_a^2 = aV_a = 9.00 & & \quad \text{(ix)} \end{aligned}$$

$$\begin{aligned} \sum x_b &= 188.50; & M_b &= 7.54; & bM_b^2 &= 1421.29; \\ \sum x_b^2 &= 1430.75; & \sum (x_b - M_b)^2 &= \sum x_b^2 - bM_b^2 = bV_b = 9.46 & & \quad \text{(x)} \end{aligned}$$

$$\frac{\sum (x_a + x_b)}{50} = 7.72 = M_{ab}; \quad nM_{ab}^2 = 2979.92;$$

$$\sum (x - M_{ab})^2 = \sum x_a^2 + \sum x_b^2 - nM_{ab}^2 = 20.08 \quad \text{(xi)}$$

$$(M_a - M_b)^2 = 0.1296 \quad \text{(xii)}$$

From (ix) and (x) we have

$$s_a^2 = \frac{\sum (x_a - M_a)^2}{n - 2} = \frac{9}{48} = 0.1875;$$

$$s_b^2 = \frac{\sum (x_b - M_b)^2}{n - 2} = \frac{9.46}{48} = 0.1971.$$

Whence from (iii):

$$t^2 = \frac{ab(M_a - M_b)^2}{n(s_a^2 + s_b^2)} = \frac{625(0.1296)}{50(0.3846)} = 4.212,$$

$$\therefore t = 2.05.$$

The corresponding ratio for the approximate  $c$ -test of 7.06 in Vol. I is given by

$$c^2 = \frac{ab(n - 1)(M_a - M_b)^2}{n \sum (x - M_{ab})^2} = \frac{625(49)(0.1296)}{50(20.08)} = 3.95$$

$$\therefore c = 1.99.$$

## 16.07 TESTING THE VARIANCE RATIO

From quite different approaches, we have arrived at the definition of two statistics respectively for testing the reality of a difference between paired sets of observations and between group means, each statistic being expressible as a ratio of independent sample variances and hence as the ratio of two independent Chi-Square variates. This ratio ( $t^2$ ) is in fact a Type VI variate, but its square root is a Type VII variate, because the numerator happens to be a Chi-Square variate of 1 d.f. Except when we have to deal with a set-up having only 2 columns and/or two rows, the variance ratios defined in 13.03-13.06 as criteria of homogeneity in connexion with the procedure known as *Analysis of Variance* involve in both dimensions Chi-Square variates with more than one degree of freedom. Usually therefore the table of the  $t$ -integral is of no assistance in assessing their significance.

On the assumption that the numerator is statistically independent of the denominator of such a ratio, the general procedure ( $F$ -test) follows from what we have learned in 15.04 and 16.04.





than unity when  $b$  is small. If the denominator of the  $F$ -ratio is  $s_b^2$  as defined in 13.03, it has the degrees of freedom corresponding to  $b$ , i.e.  $(r-1)(c-1)$ . For a  $3 \times 4$  table  $(r-1)(c-1) = 6$  and the mean value of  $F$  will be 1.5.

The probability ( $P_u$ ) that the ratio  $F$  will have a value as great as or greater than  $u$ , which we here assume to be greater than its expected value  $\mu_1$ , is given by

$$P_u = \int_u^\infty \phi(F) dF \quad \text{and} \quad \int_0^u \phi(F) dF = 1 - P_u \quad . \quad . \quad . \quad \text{(viii)}$$

Of course,  $F$  may also be less than its expected value, and the probability ( $P_v$ ) that it will be as small as or less than  $v$  is given by

$$P_v = \int_0^v \phi(F) dF \quad . \quad . \quad . \quad . \quad . \quad . \quad \text{(ix)}$$

To say that 5 per cent. of the area bounded by the curve lies in the range from  $u$  to  $\infty$ , i.e.  $P_u = 0.05$  means that the odds are about 20 : 1 against getting a value of  $F$  as great as or greater than  $u$ . Likewise  $P_v = 0.05$  signifies that the area from 0 to  $v$  is also 5 per cent. of the area bounded by the curve, and the odds are about 20 : 1 against getting a value of  $F$  no greater than  $v$ . However we need not concern ourselves with the improbability of getting a value of  $F$  less than the mean, if we take advantage of the reciprocal property of the Type VI variate (Ex. 3, 15.04). If  $Z$  is the reciprocal of  $F$ , i.e.  $Z = F^{-1}$ ,

$$f(Z) = \frac{\left(\frac{a}{b}\right)^{\frac{1}{2}a} Z^{\frac{1}{2}(b-2)}}{B(\frac{1}{2}a, \frac{1}{2}b) \left(\frac{a}{b} + Z\right)^{\frac{1}{2}(a+b)}} \quad . \quad . \quad . \quad . \quad . \quad \text{(x)}$$

This is a Type VI variate of the same form as (vi) with interchange of constants, i.e. degrees of freedom. When  $F = u$ ,  $Z = u^{-1}$  and when  $F = \infty$ ,  $Z = 0$ . The change of sign in the transformation from (vi) to (x) as we have noted in 15.02 means that

$$\begin{aligned} \int_u^\infty \phi(F) dF &= - \int_{1/u}^0 f(Z) dZ, \\ \therefore \int_u^\infty \phi(F) dF &= \int_0^{1/u} f(Z) dZ. \end{aligned}$$

If  $v = u^{-1}$ , the probability that  $F$  will have a value as great as or greater than  $u$  is therefore the same as the probability that  $Z$  will have a value as small as or less than  $v$ . Now our concern is merely with the consistency of the estimates  $s_a^2$  and  $s_b^2$ . It is therefore immaterial whether we chose  $s_a^2$  as the numerator and  $s_b^2$  as the denominator of  $F$  or *vice versa*, so long as we use the appropriate Type VI variate, *viz.* (vi) or (x), as the case may be. For economy of tabulation we may define  $s_a^2$  as the *greater* of the two estimates, so that  $F$  itself is always greater than unity.

Evidently complete tables of  $F$  so defined for a wide range of corresponding values of  $a$  and  $b$  would fill a bulky volume. Snedecor's condensed table gives two entries in each cell for corresponding values of  $a$  ( $a_1, a_2$ , etc. below) and  $b$  ( $b_1, b_2$ , etc. below). The first ( $F_{0.05}$ ) is the numerical value of the variance ratio bounding the 5 per cent. tail of the distribution, and the second ( $F_{0.01}$ ) is that which bounds the 1 per cent. tail, i.e. the odds are about 20 : 1 against getting a value of  $F$  as great as or greater than  $F_{0.05}$  and the odds are about 100 : 1 against getting a value of  $F$  as great as or greater than  $F_{0.01}$ . The lay-out is then as follows :

$\begin{smallmatrix} a \\ b \end{smallmatrix}$	$a_1$	$a_2$	$a_3$	...
$b_1$				...
$b_2$			$\begin{smallmatrix} F_{0.05} \\ F_{0.01} \end{smallmatrix}$	...
$b_3$				...
...				

If  $a = 4$ , and  $b = 8$ , the two entries are 3.84 and 7.01. This means that for the ratio  $F = s_a^2 \div s_b^2$  of two estimates  $s_a^2$  with 4 and  $s_b^2$  with 8 degrees of freedom, the odds are less than 20 : 1 against getting a value  $F > 3.84$  and less than 100 : 1 against getting a value  $F > 7.01$ . In other words,  $F = 3.5$  is below the 5 per cent. significance level,  $F = 5$  is above the 5 per cent. but below the 1 per cent. significance level,  $F = 10$  is above the 1 per cent. significance level, i.e. there are adverse odds of more than 100 : 1 against getting a value so large. Any such value of  $F$  of course implies that  $s_a^2 > s_b^2$ . The reciprocal of the value 7.01 is approximately 0.14. If our tables recorded the value of  $F$  for  $s_b^2 > s_a^2$  we should therefore find the entry  $P_{0.01} = 0.14$  for  $a = 8$  and  $b = 4$ .

*En passant*, it is worth while to recall the distinction we have drawn between *vector* and *modular* probability in Chapter 5 of Vol. I. When we speak of a 5 per cent. significance level in connexion with the normal distribution, that of the  $t$ -variate or of other symmetrical function with mean as origin, we commonly specify a range of numerically equivalent score values of *opposite* sign with a total expectation of 0.05, i.e. an expectation of 0.95 that the score value will neither be as great as a given score value nor as little as the same score value with reverse sign. This is a significance level referable to modular probability. The significance level we specify above in connexion with the  $F$  ratio is referable to vector probability as defined elsewhere, since we are concerned only with the improbability of getting a value as great as or greater than the observed one.

To justify the use of the  $F$ -test it is, of course, necessary to establish the statistical independence of the statistics  $s_a^2$  and  $s_b^2$  in the numerator and denominator. In 13.03 we have seen that homogeneity w.r.t. 2 criteria of classification implies the consistency of the estimate  $s_z^2$  with  $s_c^2$  and  $s_r^2$ , i.e.

$$F_c = \frac{(c-1)s_c^2}{(r-1)(c-1)s_z^2} \quad \text{and} \quad F_r = \frac{(r-1)s_r^2}{(r-1)(c-1)s_z^2}.$$

For the  $3 \times 4$  table we obtained in 16.04 above :

$$\frac{(r-1)(c-1)s_z^2}{\sigma^2} = u_7^2 + u_8^2 + u_9^2 + u_{10}^2 + u_{11}^2 + u_{12}^2.$$

By definition

$$\frac{(r-1)s_r^2}{\sigma^2} = \sum_{j=1}^{j=r} \frac{(M_j - M_x)^2}{\sigma_r^2} = \sum_{j=1}^{j=r} \frac{(M_j - M)^2}{\sigma_r^2} - \frac{(M_x - M)^2}{\sigma_x^2}.$$



or the same set-up, therefore,

$$\frac{(r-1)s_r^2}{\sigma^2} = (u_1^2 + u_2^2 + u_3^2) - u_1^2 = u_2^2 + u_3^2.$$

More generally,

$$\frac{(r-1)(c-1)s_z^2}{\sigma^2} = \sum_{s=r+c}^{s=rc} u_s^2 \quad \text{and} \quad \frac{(r-1)s_r^2}{\sigma^2} = \sum_{s=2}^{s=r} u_s^2.$$

Thus the denominator of  $F_r$  so expressed contains no one of the independent  $u$ -scores present in the numerator. Similar remarks apply *mutatis mutandis* to the ratio  $F_c$ .

For a rapid computation we employ the symbolism of 11.05 (p. 448-458) and 13.04 (p. 552), viz. :

$$S_a = \sum_{j=1}^{j=r} \sum_{i=1}^{i=c} x_{ij}^2; \quad S = rc \cdot M^2; \quad S_c = r \sum_{i=1}^{i=c} M_i^2; \quad S_r = c \sum_{j=1}^{j=r} M_j^2.$$

In this notation

$$(c-1)s_c^2 = S_c - S; \quad (r-1)s_r^2 = S_r - S; \quad (r-1)(c-1)s_z^2 = S_a + S - S_c - S_r.$$

Whence for testing column and row effects against residual variance,

$$F_c = \frac{s_c^2}{s_z^2} = \frac{(r-1)(S_c - S)}{S_a + S - S_c - S_r}; \quad F_r = \frac{s_r^2}{s_z^2} = \frac{(c-1)(S_r - S)}{S_a + S - S_c - S_r} \quad . \quad . \quad (xi)$$

The case which arises when there are only 2 columns, so that  $(c-1) = 1$  and hence there are  $r$  pairs of scores, is of special interest. In this case we may label the two column means as  $M_a$  and  $M_b$  respectively, distinguishing corresponding raw-scores as  $x_{aj}$  and  $x_{bj}$ , so that  $M_j = \frac{1}{2}(x_{aj} + x_{bj})$  and

$$S_a = \sum_{j=1}^{j=r} (x_{aj}^2 + x_{bj}^2); \quad S = 2r \cdot M^2 = \frac{r}{2}(M_a + M_b)^2;$$

$$S_c = r(M_a^2 + M_b^2); \quad S_r = c \sum_{j=1}^{j=r} M_j^2 = \frac{1}{2} \sum_{j=1}^{j=r} (x_{aj} + x_{bj})^2.$$

It is possible to express these quantities in terms of the paired score differences  $(x_{aj} - x_{bj}) = d_j$  and their mean  $M_d = (M_a - M_b)$ , in virtue of the identities

$$x_{aj} + x_{bj} = 2x_{aj} - d_j \quad \text{and} \quad M_a + M_b = 2M_a - M_d.$$

We then have

$$S_a = \sum_{j=1}^{j=r} (2x_{aj}^2 - 2x_{aj} \cdot d_j + d_j^2); \quad S = 2r \cdot M_a^2 - 2r \cdot M_a \cdot M_d + \frac{1}{2}r \cdot M_d^2;$$

$$S_r = \sum_{j=1}^{j=r} (2x_{aj}^2 - 2x_{aj} \cdot d_j + \frac{1}{2}d_j^2); \quad S_c = 2r \cdot M_a^2 - 2r \cdot M_a \cdot M_d + r \cdot M_d^2;$$

$$S_c - S = \frac{r}{2}M_d^2 \quad \text{and} \quad S_a + S - S_c - S_r = \frac{1}{2} \sum_{j=1}^{j=r} d_j^2 - \frac{1}{2}r \cdot M_d^2.$$

Whence by substitution in (xi) we obtain

$$F_c = \frac{r(r-1)M_d^2}{\sum_{j=1}^{j=r} d_j^2 - r \cdot M_d^2} = \frac{r(r-1)M_d^2}{\sum_{j=1}^{j=r} (d_j - M_d)^2} = t^2 \quad . \quad . \quad . \quad (xii)$$

*Numerical Example.* The following simple set of figures will serve to illustrate both computations involved in the use of the  $F$ -test to evaluate the nullity of the column criterion and its identity with the  $t$ -test for paired differences when there are only 2 columns and 2 criteria of classification. The set-up is for 3 rows (individuals) and 2 columns (successive score values) on the same individual :

	$i = 1$	$i = 2$
$j = 1$	2	4
$j = 2$	3	3
$j = 3$	4	8

Whence we have

$$\text{Column Means : } M_{1.} = 3 ; M_{2.} = 5.$$

$$\text{Row Means : } M_{.1} = 3 ; M_{.2} = 3 ; M_{.3} = 6.$$

$$\text{Grand Mean : } M = 4.$$

$$\text{Total Sum of Squares : } S_a = 118.$$

$$S = rc \cdot M^2 = 96 ; S_c = r \sum_{i=1}^{i=2} M_i^2 = 3(3^2 + 5^2) = 102.$$

$$S_r = c \sum_{j=1}^{j=3} M_j^2 = 2(3^2 + 3^2 + 6^2) = 108 ;$$

$$F = \frac{s_e^2}{s_z^2} = \frac{S_c - S}{c - 1} \frac{(r - 1)(c - 1)}{S_a + S - S_c - S_r} = \frac{2(102 - 96)}{118 + 96 - 102 - 108} = 3.$$

The differences are

$$d_1 = -2 ; d_2 = 0 ; d_3 = -4 ; M_d = -2 ;$$

$$\sum_{j=1}^{j=3} (d_j - M_d)^2 = 0 + 4 + 4 = 8 ;$$

$$t^2 = \frac{r(r-1)M_d^2}{\sum_{j=1}^{j=3} (d_j - M_d)^2} = \frac{3 \cdot 2 \cdot 4}{8} = 3.$$

\* \* \* \* \*

*Evaluation of  $t$ -Test for Paired Differences.* From one viewpoint the identity exhibited in (xii), namely, that  $F_c$  is equivalent to the so-called Student statistic  $t^2$  for  $r$  paired differences when  $c = 2$  is not surprising, inasmuch as we have seen that Type VII is the distribution of the square root of a Type VI variate when the numerator of the  $F$ -ratio is a Chi-Square variate of 1 d.f. Indeed, we obtained (vii) in 16.05 from the same equation as (iii) above. If  $f$  stands for the d.f. of the  $t$ -variate, the relevant substitutions made in deriving (vii) of 16.05 and (vi) imply the formal identity established above, viz. :

$$x^{\frac{1}{2}} = \frac{t}{\sqrt{f}} \quad \text{and} \quad x = \frac{F_c}{f} \quad \text{when} \quad c = 2, f = (r - 1).$$



In short, the paired difference test based on the  $t$ -distribution of 16.05 is exactly the same as a homogeneity test when the end in view is to decide whether a difference between column means is wholly attributable to residual sources of variation after eliminating variation associated with the row (*pairing*) effect. Now this identity brings sharply into focus a latent and not commonly recognized assumption in the prescription of the  $t$ -test for paired differences. Implicitly, we postulate that any row effect is strictly additive, in accordance with the following schema of score components :

Column A	Column B	Difference
$x_{a1} = e_{a1} + F_1$	$x_{b1} = e_{b1} + F_1$	$d_1 = e_{a1} - e_{b1}$
$x_{a2} = e_{a2} + F_2$	$x_{b2} = e_{b2} + F_2$	$d_2 = e_{a2} - e_{b2}$
$x_{a3} = e_{a3} + F_3$	$x_{b3} = e_{b3} + F_3$	$d_3 = e_{a3} - e_{b3}$
.....	.....	.....
$x_{ar} = e_{ar} + F_r$	$x_{br} = e_{br} + F_r$	$d_r = e_{ar} - e_{br}$

In this assumed schema, the dispersion of our  $e$ -components accounts for all residual variation, if the null hypothesis under consideration is correct. Hence we may regard each pair as a sample from a sub-universe which differs from any other such sub-universe only in virtue of a factor  $F$ , determining the origin of the score distribution. In other words, our latent assumption is that we take our samples from strata of a score distribution different *inter se* in one respect alone, *viz.*, that the mean score values of different strata are different, the variances of the sub-distributions being identical.

That we do, in fact, assume equality of variance in the derivation of the  $t$ -test prescribed for paired differences will be evident, if we retrace our steps to the beginning of 16.05. To express the square of the mean  $d$ -score in standardised form as an eliminable component of the denominator and hence to establish that the denominator and numerator are independent variates, we have to assume that we draw each  $d$ -score from a normal sub-universe with the same definitive parameter  $\sigma$ . Otherwise, the appropriate orthogonal transformation is unrealisable.

This raises an issue of practical importance : in what circumstances can we invoke the  $t$ -test as an appropriate procedure, if we take advantage of the possibility of pairing observations in the design of an enquiry ? The applicability of a  $t$ -test to evaluate the odds against a mean paired difference score exceeding its expected value of zero by such and such in fact demands answers to two questions : (a) whether we are entitled to regard members of one pair as different from another only in the sense that the mean of an indefinitely large number of observations on members of one pair may differ from the mean of an indefinitely large number of observations on another ; (b) whether we are entitled to regard successive observations on members of the same pair as referable to a normal universe.

To the first question we can give a positive answer in very restricted circumstances, e.g. (a) if the members of a pair constitute measurements respectively made on one and the same individual before and after some treatment procedure ; (b) if also the interval is sufficiently short to justify the assumption of no relevant change on the part of the individual and hence no source of variation other than error of measurement. The null hypothesis is then that paired differences arise in virtue of errors of measurement alone. If the technique of estimation is the same w.r.t. all such pairs, this implies that the variance of score values for successive individual measurements on one pair is the same as for all others. The example given in 7.07 (p. 315) conforms to this requirement. Each pair of observations involves local measurement of haemoglobin of the finger of one individual : (a) before ligation, (b) after ligation of the same finger of the same individual, the intervening period being short. In this set-up, the null hypothesis that ligation has no effect implies that the only source of variation is error of haemoglobin estimation.



By pairing our observations we make it possible to eliminate the variation arising from the fact that mean haemoglobin observations made on different individuals are likely to be different ; but the mere fact that measurements of a pair refer to the same individual is not always a sufficient guarantee that our sample of  $d$ -scores is referable to sub-universes of equal variance. If the interval is long we may introduce *diurnal* or other sources of variation which we cannot interpret with confidence. Such is indeed a permissible criticism of the choice of a widely quoted experiment used by Gossett himself to illustrate the uses of the  $t$ -test. In experiments on highly standardised laboratory stocks reared under a highly standardised regimen, we may be entitled to dismiss this consideration ; but pairing of observations on different individuals will rarely justify the use of a significance test based on the assumption that the  $d$ -scores are samples from a homogeneous normal universe. If we investigate the effect of an increased milk ration on the growth of children, we may eliminate certain fairly obvious sources of error by selecting as one member of a pair for treatment an individual of the same sex, age and build as the untreated member ; but we have no sufficient reason for assuming that variation of response to treatment is uniform from pair to pair.

The fact that we can rarely invoke with propriety the postulate of equal variance, unless the only residual source of variation is instrumental, reinforces the relevance of the second question raised above. In many types of experimentation, the range of sampling variation is small and the number of score values consistent therewith is also small. In short, the situations in which we can most confidently condone the assumptions implicit in the  $t$ -test for paired score differences are often such as to dictate justifiable caution w.r.t. the assumption of normality. Fortunately, our excursion into the burette universe of 14.05 has a lesson to offer in this context, *viz.*, that the mean difference of *repeated* pairs of observations on the material tends to normality fairly rapidly even if the number of different values consistent with good workmanship is small.

Since we should therefore exercise restraint in using the  $t$ -test for paired differences, it is important to realise that the  $c$ -test prescribed in 7.07 of Vol. I and referred to in 16.05 of this chapter is for reasons set forth more fully in 14.08 free from the objection that it implicitly postulates equality of variance. Subject to the recognition that it shares with any approximate  $c$ -test (including the so-called Chi-Square Test for the  $2 \times 2$  contingency table) a measure of uncertainty arising from the use of an estimate of the variance of the mean, it is equally applicable to any situation in which we can eliminate known sources of variation whether by pairing observations on the same individual or by pairing like individuals, as when one compensates for unequal overhead illumination or local temperature differences due to draughts by pairing off in a greenhouse pots of plants of the same stock subjected to different treatments.

#### 16.08 THE CORRELATION RATIO

Any significance test for homogeneity derives its rationale from the distribution of *independent* estimates of the variance ( $\sigma^2$ ) of the unit sample score distribution of the putative common universe. For a set-up involving one criterion of classification associated with the columns, our statistics are respectively referable to the variance of the column-means and of the mean variance of scores within the columns. For a 2-way lay-out we employ one estimate referable either to the column-means or to the row-means and a second statistic ( $s_t^2$ ) referable to the difference between the total variance ( $V_x$ ) and the sum of these two. At first sight it would seem more reasonable to choose an estimate based on the total variance as our second statistic in either case, i.e.  $s_t^2$  defined in 16.04 as

$$s_t^2 = \frac{rc}{(rc - 1)} V_x \quad . \quad . \quad . \quad . \quad . \quad . \quad (i)$$



From this expression we can derive the Chi-Square variate with  $(rc - 1)$  degrees of freedom :

$$\frac{(rc-1)s_i^2}{\sigma^2} \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad (\text{ii})$$

Thus  $z$  defined as below defines the ratio of 2 Chi-Square variates, involving the ratio of 2 unbiased estimates of  $\sigma^2$ :

$$z = \frac{(c-1)s_c^2}{(rc-1)s_r^2} \quad . \quad . \quad . \quad . \quad . \quad . \quad (\text{iii})$$

If we make the appropriate orthogonal transformation we see that this is *not* the ratio of 2 independent Chi-Square variates. Thus we may express  $z$  in accordance with the procedure of 16.04 as

$$\begin{aligned} \mathfrak{s} &= \frac{u_2^2 + u_3^2 + u_4^2 \dots u_c^2}{u_2^2 + u_3^2 + u_4^2 \dots u_{rc}^2} \quad . \quad . \quad . \quad . \quad . \quad . \quad (\text{iv}) \\ &= \frac{u_2^2 + u_3^2 \dots u_c^2}{(u_2^2 + u_3^2 \dots u_c^2) + (u_{c+1}^2 + u_{c+2}^2 \dots u_{rc}^2)}. \end{aligned}$$

In the foregoing expression the denominator contains all the square  $u$ -scores present in the numerator, but we can express it as a function of the ratio of two independent sets of  $u$ -scores if we divide the numerator and denominator by  $(u_2^2 + u_3^2 \dots u_c^2)$ , so that

$$z = \frac{1}{1 + \frac{u_{c+1}^2 + u_{c+2}^2 \dots u_{rc}^2}{u_2^2 + u_3^2 \dots u_c^2}} = \frac{1}{1+x} \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad (v)$$

In (v) the numerator of  $x$  contains  $rc - (c + 1) + 1 = c(r - 1)$  terms like  $s_d^2$  of (xxvii) in 16.04, if the columns contain an equal number ( $r$ ) of rows. When our concern is with only one criterion of classification, we may write  $rc = n$  as in (xxviii) of 16.04, so that  $c(r - 1) = (n - c)$ . The denominator contains  $(c - 1)$  independent terms, so that  $x$  has the same distribution as

$$\frac{(n - c)s_d^2}{(c - 1)s_c^2}.$$

We might arrive at this conclusion by using the tautology of the grid. By definition,

$$s_c^2 = \frac{rc}{c-1} V(M_c) = \frac{n}{c-1} V(M_c) \quad \text{and} \quad s_i^2 = \frac{rc}{(rc-1)} V_x = \frac{n}{(n-1)} V_x,$$

$$\therefore \frac{c-1}{n}s_c^2 = V(M_c) \quad \text{and} \quad \frac{n-1}{n}s_t^2 = V_x = V(M_c) + M(V_c),$$

$$\therefore \alpha = \frac{V(M_c)}{V(M_c) + M(V_c)} = \frac{1}{1 + \frac{M(V_c)}{V(M_c)}}.$$

In the above

$$\frac{(n - c)}{n} s_d^2 = M(V_c),$$

$$\therefore \frac{M(V_c)}{V(M_c)} = \frac{(n-c)s_d^2}{(c-1)s_c^2} = x \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \text{(vi)}$$

Thus  $x$  is the ratio of two independent Chi-Square variates. If we write

$$a = (n - c) \quad \text{and} \quad b = (c - 1)$$

for brevity, so that  $a$  and  $b$  are the d.f. of numerator and denominator, the p.d. equation of  $x$  is given by

$$f(x) = \frac{x^{\frac{1}{2}(a-2)}}{B(\frac{1}{2}a, \frac{1}{2}b) \cdot (1+x)^{\frac{1}{2}(a+b)}} \quad \text{. . . . . (vii)}$$

In 15.02 we have seen that the p.d. equation of  $z = (1+x)^{-1}$  is therefore

$$F(z) = \frac{z^{\frac{1}{2}(b-2)} (1-z)^{\frac{1}{2}(a-2)}}{B(\frac{1}{2}a, \frac{1}{2}b)}.$$

Or more fully

$$F(z) = \frac{z^{\frac{1}{2}(c-3)} (1-z)^{\frac{1}{2}(n-c-2)}}{B\left(\frac{c-1}{2}, \frac{n-c}{2}\right)} \quad \text{. . . . . (viii)}$$

Thus the distribution of the ratio defined by  $z$  is a Type I variate within the framework of the implicit assumption that the column samples come from *the same normal universe*.

Evidently the tabulation of the Type I variate  $z$  could give us no information which we cannot derive from the Type VI variate  $x$  if our only concern were to establish homogeneity w.r.t. the column criterion of classification. The interest of its distribution lies in the fact that we can express either *correlation* ratio of a bivariate distribution as a function of the same form. We have seen in Chapter 10 of Vol. I that it is always possible to lay-out a bivariate frequency grid as a grid of one or other set of score values, as in the numerical example of the accompanying table; and we may analyse either set of scores w.r.t. one criterion of classification by putting in the same column scores which go with one and the same value of the alternative set. The number of rows in each column will not necessarily be the same; but we have seen (13.07) that this is immaterial, when our concern is with a single taxonomic criterion, here taken as that of the column heading. The 2 score-grids so constructed respectively set out the relevant data for the evaluation of

$$\eta_{ab}^2 = \frac{V(M_{ab})}{V_a} \quad \text{and} \quad \eta_{ba}^2 = \frac{V(M_{ba})}{V_b} \quad \text{. . . . . (ix)}$$

Frequency Grid

		A-scores				Total No.	$M_{ab}$
		0	1	2	3		
B-scores	0	1	2	1	0	4	1
	1	1	1	2	1	5	$\frac{8}{5}$
	2	0	2	2	3	7	$\frac{15}{7}$
	Total No.	2	5	5	4	16	$M_a = \frac{27}{16}$
$M_{ba}$		1	5	$\frac{6}{5}$	$\frac{7}{4}$	$M_b = \frac{5}{4}$	



*Score-Grid (A-scores)*

	$C_1$ ( $b = 0$ )	$C_2$ ( $b = 1$ )	$C_3$ ( $b = 2$ )	
	0	0	1	
	1	1	1	
	1	2	2	
	2	2	2	
	..	3	3	
	..	..	3	
	..	..	3	
				<i>Total</i>
$r_t$	4	5	7	16
$M_t = M_{ab}$	1	$\frac{8}{5}$	$\frac{15}{7}$	$M_a = \frac{27}{16}$

*Score-Grid (B-scores)*

	$C_1$ ( $a = 0$ )	$C_2$ ( $a = 1$ )	$C_3$ ( $a = 2$ )	$C_4$ ( $a = 3$ )	
	0	0	0	1	
	1	0	1	2	
	..	1	1	2	
	..	2	2	2	
	..	2	2	..	
					<i>Total</i>
$r_t$	2	5	5	4	16
$M_t = M_{ba}$	$\frac{1}{2}$	$\frac{5}{5}$	$\frac{6}{5}$	$\frac{7}{4}$	$M_b = \frac{5}{4}$

In the symbolism of this context we may define either ratio with appropriate alternative interpretation of the column criterion as

$$\eta^2 = \frac{V(M_c)}{V_x} \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \quad (\text{x})$$

In this expression  $V_x$  is the total score distribution. If the grid items are  $A$ -scores,  $V_x = V_a$ ; and  $M_c$ , being the mean  $A$ -score for a column defined by a fixed  $B$ -score value,  $M_c = V_{ab}$ . We then interpret the above  $\eta_{ab}^2$  in accordance with (ix). Conversely, the  $B$ -score-grid  $V_x = V_b$ ,  $M_c = M_{ba}$  and  $\eta^2 = \eta_{ba}^2$ . For either grid, the total number of entries is  $n = rc$  if the columns contain equal number of items, otherwise

$$n = c \sum_{i=1}^{i=c} r_i.$$

In either case, we may put

$$\frac{(n-1)s_t^2}{n} = V_x \quad \text{and} \quad \frac{(c-1)s_c^2}{n} = V(M_c),$$

$$\therefore \frac{(c-1)s_c^2}{(n-1)s_t^2} = \eta^2.$$





$$V(M_c) = \sum_{i=1}^{i=4} \frac{r_i(M_i - M_b)^2}{n} = \frac{2}{16}(\frac{1}{2} - \frac{5}{4})^2 + \frac{5}{16}(1 - \frac{5}{4})^2 + \frac{5}{16}(\frac{6}{5} - \frac{5}{4})^2 + \frac{4}{16}(\frac{7}{4} - \frac{5}{4})^2 = 0.1531;$$

$$V_x = \frac{2(0 - \frac{5}{4})^2 + 5(1 - \frac{5}{4})^2 + 5(2 - \frac{5}{4})^2 + 4(3 - \frac{5}{4})^2}{16} = 1.1563;$$

$$\eta_{ba}^2 = \frac{V(M_c)}{V_x} = 0.1324;$$

$$\frac{\eta_{ba}^2}{1 - \eta_{ba}^2} = \frac{0.1324}{0.8676} = 0.1526;$$

$$F = \frac{(16 - 4)}{(4 - 1)} \cdot \frac{\eta_{ba}}{1 - \eta_{ba}^2} = 0.6104.$$

In this case  $F < 1$ . Since the  $F$ -table cites significance levels only for values exceeding expectation, we make use of the reciprocal property of the  $F$ -variate, i.e. we ask what would be the expectation of getting a value as great as  $(0.6104)^{-1} = 1.54$ . We must then reverse the degrees of freedom, by putting  $16 - 4 = a = 12$  and  $(4 - 1) = b = 3$ . The  $F$ -table gives

$$a = 12, b = 3$$

$$\begin{array}{ll} 5 \text{ per cent. level} & . \quad 8.74 \end{array}$$

$$\begin{array}{ll} 1 \text{ per cent. level} & . \quad 27.05 \end{array}$$

Again the result shows no significant departure from zero correlation.

## CHAPTER 17

# REGRESSION AND DISCRIMINATION

### 17.00 REGRESSION IN REAL WORK

So far we have gained our acquaintance with the concept of regression only in the domain of statistical models based on games of chance. We shall now consider it as a tool in the day's work. Though the statistical procedure subsumed under the term *regression* owes its name to Galton and its literal meaning to Galton's erroneous views about inheritance involving many gene substitutions, it is essentially one which physicists have used under a different designation for more than a century as a curve-fitting device due to Gauss. It is not easy to evaluate its proper uses nor to recognise what pitfalls beset its applications unless we delve into this background, as we shall now do.

In the domain of statistical models, it suffices to define regression of the linear type in purely algebric terms, *viz.* regression of the *B*-score on the *A*-score is linear if the *B*-score means respectively associated with successive equally spaced values of the *A*-score increase by equal increments. To say this is to say that they constitute an arithmetic series and as such would fall on the same straight line if plotted against the *A*-scores graphically. When discussing experimental data, the geometrical definition has certain advantages which will emerge in what follows.

The Gaussian origins of regression have to do with the problem of determining agreed values of the definitive constants of a *straight line law*. Needless to say, a linear law is not a common type in the exact sciences; but a suitable score transformation suffices to reduce a non-periodic physical law to the linear form. For instance, we can express Boyle's law ( $p v = k$ ) in the form  $p = k d$  by the substitution  $d = v^{-1}$ , and plot mean values of  $d$  against fixed values of  $p$  to determine the regression constant  $k$ . Thus the issue, as stated above, is of more general interest than would appear at first sight. It arises in a multiplicity of situations which demand the adoption of a universally accepted numerical value for a physical constant *on the implicit assumption that there exists a true law of a type amenable to expression in linear form*.

It is of paramount importance to recognise the implications of the assumption last stated. It may be easier to get them into focus if we take the simplest possible example of a physical law as a type specimen. One of the few examples of a familiar physical law commonly stated in the linear form is the law of Hooke (*ut tensio sic vis*) connecting the length ( $l$ ) of a spring with the load or tension ( $t$ ) applied for values of the latter not too near the elastic limit. The law is linear if

$$l = k \cdot t + C.$$

In textbooks we commonly meet it in the form which the above assumes, if we denote by  $l_0$  the unstretched length when  $t = 0$ :

$$(l - l_0) = k \cdot t.$$

So stated,  $(l - l_0) = s$  is the stretch, and we may more briefly express the law as  $s = kt$ . This is, however, an idealised statement of any laboratory situation.

In real life we do not expect that all our observations, however carefully made, will fall exactly on a straight line or other descriptive curve. Even if we can eliminate all extraneous sources of variation, e.g. condensation of moisture on the scale pan, we know that successive observations involving no change of the controlled variable—as when we use the same box of weights—will involve instrumental errors, e.g. discrepancies in successive readings of a vernier scale; but we may have reason to believe that errors in this sense are not systematic or at least



that we can arrange matters so that they are not. This is to say that numerically equivalent positive and negative deviations about an assumed true value cancel one another in the long run, the mean value of successive observations referable to one and the same value of the variable under experimental control—in this case *load*—being therefore an unbiased estimate of the one deemed to be *true* in this sense.

In so far as we can rightly make this assumption, sufficiently justified by experience in many laboratory situations, statement of the law is wholly explicit if we write  $M_{s,t}$  as the mean of an indefinitely large number of  $s$ -scores referable to the same  $t$ -score, so that

$$M_{s.t} = k.t.$$

For a fixed set of  $t$ -values we may write  $E(t) = M_t$  and  $E(M_{s..t}) = M_s$ , so that

$$M_{s+t} - M_s = k(t - M_t) \quad . \quad . \quad . \quad . \quad . \quad (i)$$

For a discrete distribution this expresses the belief that the universe or true mean value of  $s$ -scores associated with particular  $t$ -scores constitute an A.P. when arranged in accordance with corresponding equally spaced successive values of the  $t$ -scores, in which case certain tautologies established in 11.04 are necessarily true. In particular

$$k = \frac{\text{Cov}(s, t)}{V_s} \quad . \quad . \quad . \quad . \quad . \quad . \quad (\text{ii})$$

Equations (i) and (ii) above define the properties of the universe of our observations. We shall later see (17.01 below) that the sample statistic calculated on the same basis as (ii) is in fact an unbiased estimate of the true value ( $k$ ) of the physical constant. First, however, it is important to be clear about all the assumptions we have so far made. Our interpretation of a physical law exhibiting the dependence of a  $B$ -score on an  $A$ -score presupposes that

- (a) to each observation  $x_{b \cdot a}$  for a fixed value of the  $A$ -score corresponds a true value—the regression score  $x_{r \cdot a}$ —from which the observation deviates by an error or  $\epsilon$ -score ( $\epsilon_{b \cdot a}$ );
- (b) the regression scores lie on a straight line each such score being identical with the corresponding mean  $B$ -score ( $M_{b \cdot a}$ ) in the universe of *all* samples;
- (c) the observed  $B$ -score ( $x_{b \cdot a}$ ) for a fixed  $A$ -score thus consists of two additive components which we may express by the relations

$$x_{b..a} = x_{r..a} + \epsilon_{b..a} \quad \text{and} \quad x_{r..a} = M_{b..a} \quad . \quad . \quad . \quad (\text{iii})$$

- (d) these components are statistically independent, the  $\epsilon$ -scores being distributed randomly about zero mean ;
- (e) the distributions of the  $B$ -score for different fixed values of the  $A$ -score therefore differ solely in virtue of the fact that different values of  $x_{r..a} = M_{b..a}$  fix the origin of the distribution, whence  $V_{b..a}$  is constant and the universe is homoscedastic in one dimension ;
- (f) since  $x_{r..a}$  is constant for a fixed value of the  $A$ -score, the corresponding variance of the  $B$ -score distribution is the variance of the  $\epsilon$ -scores, i.e.

$$V_{\bar{p}, \alpha} = V(\epsilon_{\bar{p}, \alpha}) \quad . \quad . \quad . \quad . \quad . \quad . \quad (\text{iv})$$

Within the sample, regression will not be exactly linear, but we can define a set of hypothetical regression scores ( $x_{r. as}$ ) having this property, as in 11.04, and a sample statistic  $k_s$  as the slope of the line on which they lie. The use of a sample statistic  $k_s$  based on corresponding sample values of  $s$  and  $t$  to estimate  $k$  in (ii), as illustrated in the numerical example at the beginning of 18.02, is what physicists call line-fitting by the *method of least squares*. We may regard the latter as a procedure for obtaining an estimate with a confidence interval as small as possible at a



particular confidence level; but a general proof that such an estimate is unbiased in the sense that the long-run mean result of applying the prescription is the true value of  $k$  is laborious. Accordingly, we shall defer its consideration till we have determined the unbiased estimate of the regression coefficient by another method. It will then take its place as a particular illustration of the principle of *minimal variance*, when we ask whether the unbiased estimate under consideration is also an *efficient* one.

The least square method derives its rationale in part from the assumption of a *normally* distributed set of  $B$ -score values and a more general approach to the issue stated at the beginning of the last paragraph is easy to visualise. In determining an unbiased estimate of  $k$ , we shall make no assumptions about the law of the distribution of errors, any such assumption being indeed highly questionable in many real situations. The reader who has consulted treatises which present the use of regression analysis against the background of the so-called *bivariate normal universe* will notice that certain assumptions commonly presented as deductions from its properties are logically implicit in the concept of a physical law, in particular for the reason stated in (e) above, *homoscedasticity*, i.e. equal variance of the  $B$ -score distribution for different values of the  $A$ -score.

Now this property is one-dimensional. Our formulation of a law relating the mean  $B$ -score to the  $A$ -score implies nothing about the distribution of the  $A$ -scores for fixed values of the  $B$ -score in the context of the experiment. The range of  $A$ -score values, and the number of each, depends on the way the investigator carries out the experiment; and we are entitled to regard any one experiment as a sample of an indefinitely large number of experiments carried out in the same way. Commonly, the fixed  $A$ -scores, or independent variable in the Cartesian sense, represent the one easiest to control; but we may often reverse the procedure. For instance, we may fix the vernier of a micrometer to read when the stretch of a spring attains a certain limit and measure repeatedly what we have to add to the scale pan to achieve the result. If so, we make  $t$  in the relation  $s = kt$  our dependent variable and must reinterpret  $k$  accordingly.

We shall refer again to this duality of regression in 18.03 below. Here it is admissible to forestall an unnecessary difficulty which the reader may have experienced if already acquainted with the concept of a *bivariate normal universe*. Within the framework of the dubious assumption that errors involved in determining the  $B$ -score for a fixed  $A$ -score and *vice versa* are both distributed normally, the solid model of a universe which is normal in both the  $B$ -dimension and the  $A$ -dimension of the grid corresponds to reality only in the sense that it embodies the possibility of carrying out an experiment in one of two ways. The conduct of any actual experiment is *ipso facto* unique in the sense that it refers to one set of fixed scores, the distribution of which depends on the observer's choice. In so far as our concern is with the uncertainties arising from error in the procedure we do in fact adopt, we cannot therefore conceive of the experiment as a sample from a universe in which the score distribution is normal in both dimensions. If we choose to take equal numbers of measurements ( $B$ -scores for each of a particular set of  $A$ -scores), the assumption of normally distributed errors implies that our sampling universe is normal in the  $B$ -dimension and rectangular in the  $A$ -dimension, and if we chose to describe it in the language of 3-dimensional geometry the frequency surface is less like the sugar loaf of the bivariate normal universe than the outside of a Nissen hut.

#### 17.01 PRINCIPLE OF THE FIXED- $A$ SET

The models of 12.01–12.09 define the unit sampling distribution of different universes of correlation. We shall now attempt to break down the distribution for samples of more than one item from a bivariate universe so conceived in accordance with the considerations advanced in 12.00.



## 2 FACE PACK - UNIT SAMPLE DISTRIBUTION

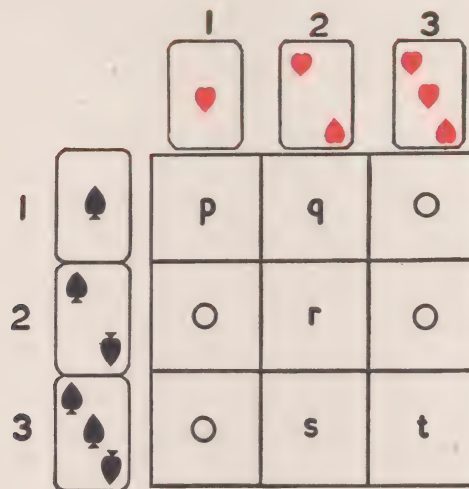
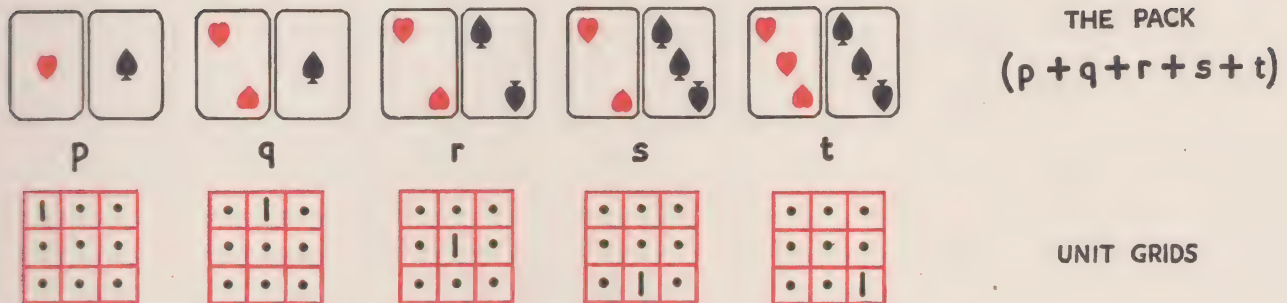


FIG. 118. For explanation see text.

Our model universe, from which we sample *with* replacement in conformity with the assumption that it is both discrete and infinite, will be a pack of cards made by gluing pairs of cards face upwards, so that one face bears 1, 2 or 3 hearts (*A*-score) and the other bears 1, 2 or 3 spades (*B*-score) as shown in Fig. 118. The universe (unit sample) bivariate distribution is

$x_a, x_b$	1.1	2.1	2.2	2.3	3.3
proportions	$p$	$q$	$r$	$s$	$t$

That we sample with replacement for the reason set forth in 12.00 means that the distribution of 2-fold samples is deducible by the chessboard procedure (Fig. 119) of which the equivalent definitive multinomial is  $(p + q + r + s + t)^2$ . By successive application of the chessboard device, we can visualise the extraction of  $n$ -fold samples in accordance with terms of the expansion of  $(p + q + r + s + t)^n$ . Thus the probability of getting one paired score of (1.1) and 2 paired scores of (2.2) in a sample of 3 is  $3pr^2$ , and the 3-fold sample grid is then

DERIVATION OF A 2-FOLD SAMPLE DISTRIBUTION  
FROM THE 2 FACE CARD PACK UNIVERSE

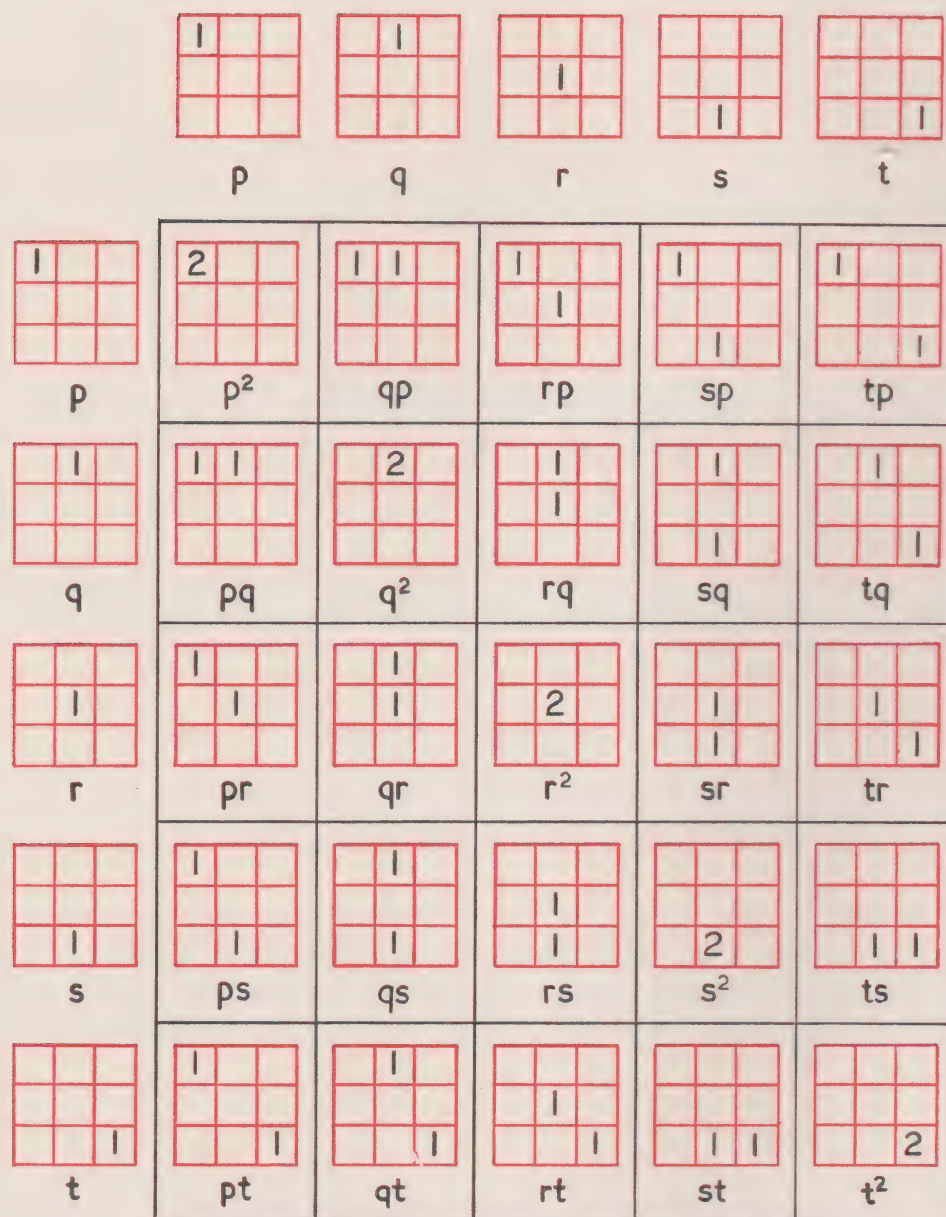


FIG. 119.



	1	2	3
1	1	0	0
2	0	2	0
3	0	0	0

$$M_a = \frac{5}{3} = M_b.$$

$$\text{Cov}(x_a, x_b) = \frac{2}{9} = V_a = V_b.$$

$$r_{ab} = 1.$$

We may classify our samples of 2, 3, etc., in *fixed-A* sets, i.e. sets of samples having the same border *A*-score distribution. For 2-fold samples (Fig. 120) of our 2-face card-pack model of 5 paired score classes, there are 6 such sets, labelled  $A_{200}$  ( $x_a = 1$  twice),  $A_{101}$  ( $x_a = 1$  once and  $x_a = 3$  once), etc. Fig. 121 shows all possible samples of 3 classified in the 10 fixed-*A* sets,  $A_{300}$ ,  $A_{210}$ ,  $A_{201}$ , etc. For the theory of sampling from a bivariate universe, a very important result is an immediate consequence of the fact that the universe of the unit sample distribution is infinite, i.e. that we may regard the replacement condition as valid. The pooled mean value of the *B*-score associated with any *A*-score present in the subset is the same for any fixed-*A* set as for the unit sample distribution referable to all permissible values of *A*. This follows from the way in which we weight the samples in conformity with the principle of equipartition of opportunity implicit in the chessboard procedure; but the student may find it instructive to check the rule by recourse to the 2-fold or 3-fold sample distribution of our 2-face card pack model as below.

For the unit sample distribution we derive the mean *B*-score ( $M_{b.2}$ ) associated with  $x_a = 2$  as follows :

$$M_{b.2} = \frac{q(1) + r(2) + s(3)}{q + r + s} = \frac{q + 2r + 3s}{q + r + s} \quad \dots \quad (i)$$

For the *fixed-A* set  $A_{120}$  we have 6 different types of sample structure of which one with frequency  $6pqr$  is

1	1	0
0	1	0
0	0	0

Within this sample the value of  $M_{b.2}$  is  $\frac{1}{2}(1) + \frac{1}{2}(2) = \frac{3}{2}$ . The sample itself consists of equal numbers of paired scores (1.1), (2.1), (2.2), with a total frequency of  $6pqr$  as stated, whence the frequency of paired scores having the relevant values  $x_a = 2$  and  $x_b = 1$  or 2 is  $4pqr$ , and this, divided by the total of such, must be our sample weight, when we pool the values of  $M_{b.2}$  for the whole set. We then summarise the computation of  $M_{b.2}$  for the entire set  $A_{120}$  as follows :

FIXED A-SET  
HISTOGRAM FOR 2 FOLD SAMPLE

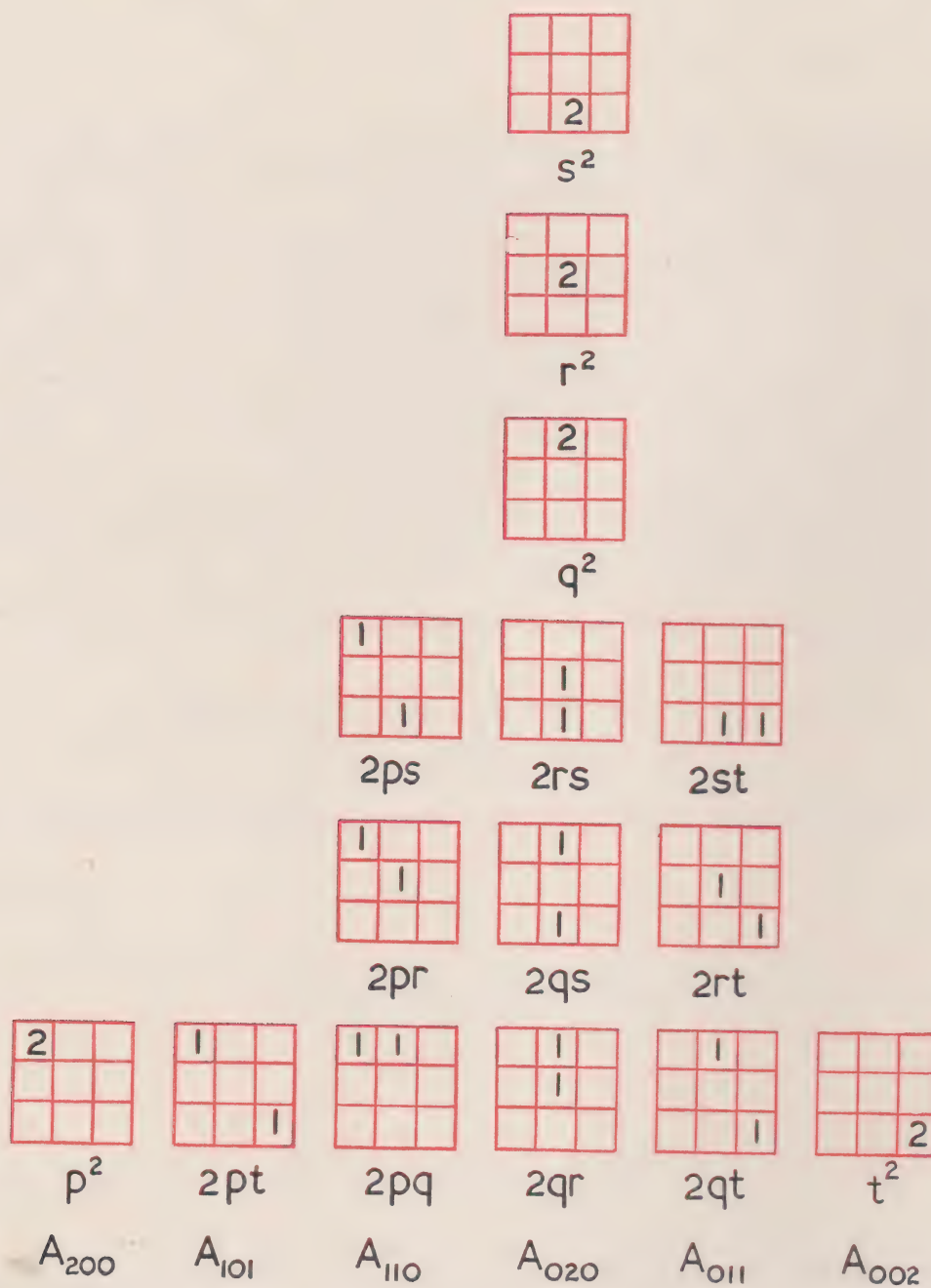


FIG. 120.



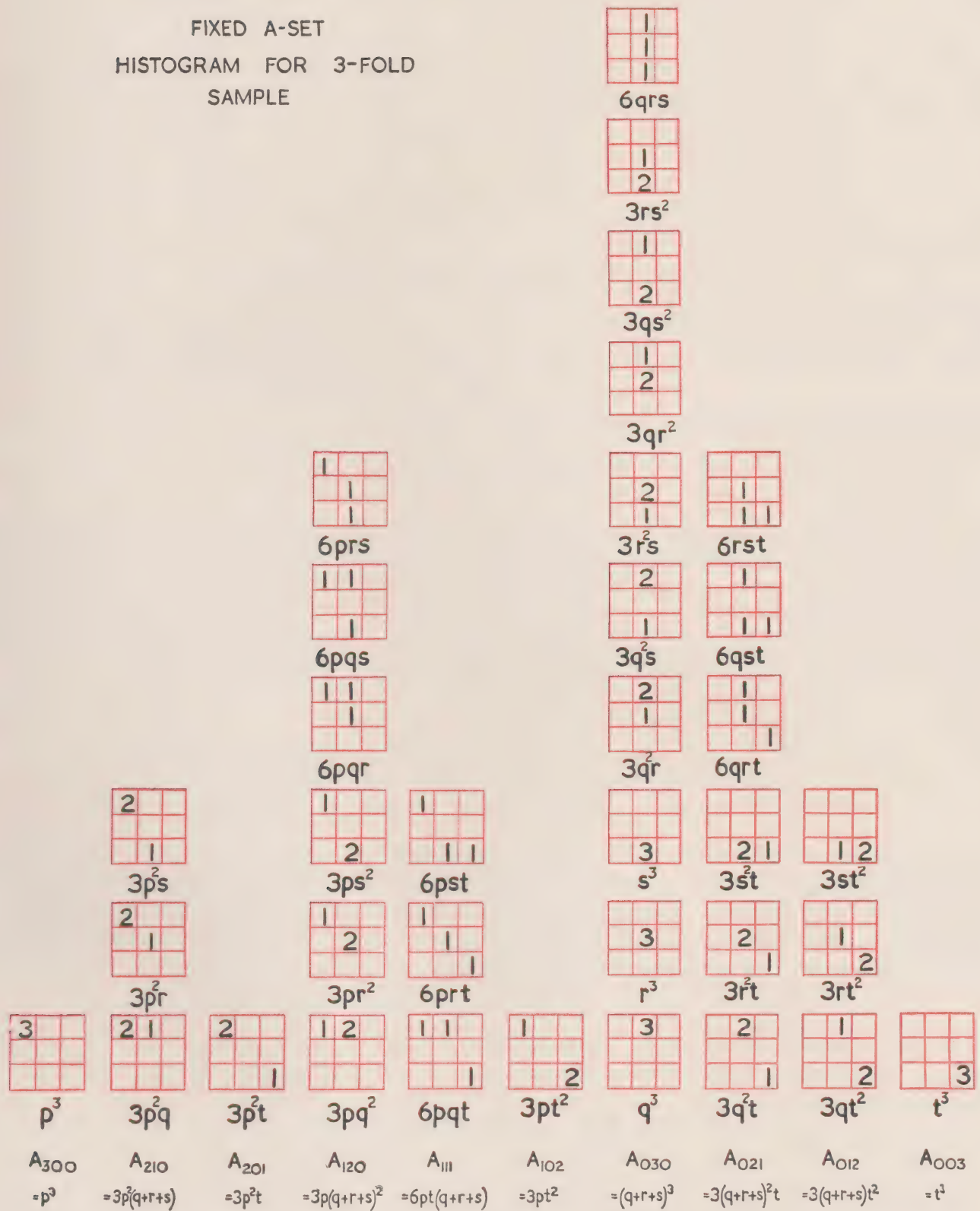


FIG. 121.

Sample Structure	Sample Frequency	$M_{b..2}$	Total Frequency of $(2 \cdot x_b)$
(1.1) . (2.1) . (2.2)	$6pqr$	$\frac{3}{2}$	$4pqr$
(1.1) . (2.1) . (2.3)	$6pqs$	2	$4pqs$
(1.1) . (2.2) . (2.3)	$6prs$	$\frac{5}{2}$	$4prs$
(1.1) . (2.1) . (2.1)	$3pq^2$	1	$2pq^2$
(1.1) . (2.2) . (2.2)	$3pr^2$	2	$2pr^2$
(1.1) . (2.3) . (2.3)	$3ps^2$	3	$2ps^2$

The total of the last column in the above is

$$2p(s^2 + r^2 + q^2 + 2rs + 2qs + 2qr) = 2p(s + r + q)^2.$$

The total of the products of the last two columns is

$$\begin{aligned} 8pqs + 10prs + 6pqr + 2pq^2 + 4pr^2 + 6ps^2 &= 2p(4qs + 5rs + 3qr + q^2 + 2r^2 + 3s^2) \\ &= 2p(s + r + q)(q + 2r + 3s). \end{aligned}$$

Whence the mean value of  $M_{b..2}$  for the entire set  $A_{120}$  is

$$\frac{2p(s + r + q)(q + 2r + 3s)}{2p(s + r + q)^2} = \frac{(q + 2r + 3s)}{s + r + q}.$$

This identity arises from a more general relation inherent in our visualisation of sampling in a bivariate universe as equivalent to sampling independently with replacement from a universe of unit grids in accordance with the chessboard principle. The *principle of the fixed set*, as we may call it, is that the distribution of  $B$ -scores for a particular value of the  $A$ -score is the same in any fixed- $A$  set and hence in the universe itself. One illustration will suffice to make this clear. From Fig. 121 we see that the  $B$ -scores for  $x_a = 2$  in the set  $A_{021}$  is

$x_{b..2}$  Weighted frequency

$$\begin{aligned} 1 & (0 + 2qst + 2qrt + 0 + 0 + 2q^2t) \div 2t(r + q + s) = q. \\ 2 & (2rst + 0 + 2qrt + 0 + 2r^2t + 0) \div 2t(r + q + s) = r. \\ 3 & (2rst + 2qst + 0 + 2s^2t + 0 + 0) \div 2t(r + q + s) = s. \end{aligned}$$

In the notation of our club-sandwich universe of all samples the principle of the fixed set is the formal identity  $M_{b..a} = M_{b..as}$ ; and its derivation follows directly from the initial assumptions and definitions of symbols. Our hypothesis is that an observed  $B$ -score ( $x_{b..a}$ ) has two *additive* components: the true value ( $M_{b..a}$ ), being the long-run mean of repeated observations referable to a fixed  $A$ -score, and a random error ( $\epsilon_{b..ac}$ ) independent thereof, so that

$$x_{b..a} = M_{b..a} + \epsilon_{b..ac} \quad \text{and} \quad M_{b..acs} = E_b(x_{ba}) = M_{b..a} + M_{e..ac}.$$

If we here conceive each stratum of our stratified universe to accommodate a sufficiently large number of experiments referable to the same set of  $A$ -scores, we presume that negative and positive errors cancel, i.e.

$$E_c \cdot E_b(\epsilon_{b..ac}) = 0 = E_c(M_{e..ac}),$$

$$\therefore M_{b..as} = E_c(M_{b..acs}) = M_{b..a} + E_c(M_{e..ac}) = M_{b..a}.$$

The model of Fig. 118 is consistent with linear regression if we put  $q = s$  in the u.s.d.; but it does not embody 2 essential properties of the Gaussian bivariate universe as defined in 17.00, viz.: (a) *homoscedasticity* or equal variance of  $B$ -scores for any fixed value of the  $A$ -score; (b) distribution of *errors* (i.e. deviations from column means) about zero mean for each fixed value of the  $A$ -score. To satisfy either condition we must have at least 2 non-zero cell entries in each



column of the universe grid, since the column variance will be zero if there is only one non-zero entry therein. The arrangement below shows how it is possible to satisfy both conditions, while still leaving two corner cells empty :

	1	2	3
1	$p$	$q$	.
2	$p$	$6q$	$t$
3	.	$q$	$t$
$M_{b..a}$	1.5	2.0	2.5
$V_{b..a}$	0.25	0.25	0.25

$$(2p + 8q + 2t) = 1.$$

$$k_{ba} = \frac{1}{2}; \quad r_{ab}^2 = \frac{(p+t) - 2(t-p)^2}{\frac{1}{2} + (p+t) - 2(t-p)^2}.$$

If the bivariate universe is homoscedastic as for the grid last shown, the principle of the fixed- $A$  set carries an important consequence. Since the distribution of  $B$ -scores for a particular value of  $A$  is the same in any fixed- $A$  set, the variance of the  $B$ -scores for a particular value of  $A$  is the same in each fixed- $A$  set and hence the same as in the universe. Hence *each fixed- $A$  set is also homoscedastic if the universe itself is homoscedastic both with respect to  $B$ -scores and to their error components*. For a particular value of the  $A$ -score the error variance is, of course, equivalent to the  $B$ -score variance, since the two distributions differ with respect to the mean only. Thus the error variance of a fixed- $A$  set as a whole, or for a particular  $A$ -score of the fixed- $A$  set is that ( $\sigma_b^2$ ) of the whole universe.

The principle of the fixed- $A$  set invites us to visualise the complete distribution of  $r$ -fold samples from a bivariate universe by conceiving it as a 3-dimensional stratified grid. Each layer specifies a particular type of sample structure in conformity with the assumption that the sample consists of any combination of  $r$  paired scores not necessarily different. Each stratum consists of all layers of a fixed- $A$  set and of no others. Within the stratum the number of identical layers tallies with the relative frequencies of the corresponding types of sample. Within the universal grid consisting of strata corresponding to every possible fixed- $A$  set, the strata repeat themselves, the numbers of strata of one or other type being proportional to the frequency of the corresponding fixed- $A$  set in the sample distribution. Having weighted our layers and strata in this way, it follows that each paired score occurs in the universal grid in the same proportion as in the unit sample distribution, whence any parameter of the whole grid is equivalent to the corresponding parameter of the *u.s.d.*

Having so conceived the universe of the  $r$ -fold sample, we may label parameters referred to below with due regard to the conclusions stated above, *viz.*: (a) that the border  $A$ -score distribution of all layers in a stratum is the same; (b) that the mean  $B$ -score for a particular value of the  $A$ -score is the same for the whole stratum as for the universe.

	Layer (sample)	Stratum (fixed- $A$ set)	Universe
Mean $B$ -score . . .	$M_{b..cs}$	$M_{b..s}$	$M_b$
Mean for fixed- $A$ set . .	$M_{b..acs}$	$M_{b..as} = M_{b..a}$	$M_{b..a} = M_{b..a}$
Variance of $B$ -score distribution . . .	$V_{b..cs}$	$V_{b..s}$	$\sigma_b^2$
Mean $A$ -score . . .	$M_{a..cs} = M_{a..s}$	$M_{a..s} = M_{a..cs}$	$M_a$
Variance of $A$ -score distribution . . .	$V_{a..cs} = V_{a..s}$	$V_{a..s} = V_{a..cs}$	$\sigma_a^2$
Covariance . . .	$Cov(x_{a..cs}, x_{b..cs})$	$Cov(x_{a..s}, x_{b..s})$	$Cov(x_a, x_b)$
Regression Coefficient .	$k_{ba..cs}$	$k_{ba..s}$	$k_{ba}$

In this set-up we can visualise the layers as frequency grids of  $a$  columns and  $b$  rows or as score grids of 2 columns, one for  $A$ -scores and one for  $B$ -scores paired row by row as in a computing schema (see schemata at the end of 11.01). For the present, we shall adopt the former convention, defining a *sequence* of operations in accordance with the conventions of 11.06 as follows :

	Within	For all values of $A$ -score	For fixed values of $A$ -score
Column . . .		$E_{b \cdot a}(\dots)$	$E_{b \cdot a}(\dots)$
Layer . . .		$E_a \cdot E_{b \cdot a}(\dots)$	$E_{b \cdot a}(\dots)$
Stratum . . .		$E_c \cdot E_a \cdot E_{b \cdot a}(\dots) \equiv$ $E_a \cdot E_c \cdot E_{b \cdot a}(\dots)$	$E_c \cdot E_{b \cdot a}(\dots)$
Whole Grid . . .		$E_s \cdot E_c \cdot E_a \cdot E_{b \cdot a}(\dots) \equiv$ $E_s \cdot E_a \cdot E_c \cdot E_{b \cdot a}(\dots)$	$E_s \cdot E_c \cdot E_{b \cdot a}(\dots)$

The order of operations is *not* interchangeable, except as indicated. Otherwise, the symbol  $E_a$  would be ambiguous, if we employed it in operations which we might otherwise distinguish as  $E_a$  referable to all  $A$ -scores and  $E_a$  referable to a sample or a fixed- $A$  set. Without ambiguity we can drop the subscripts  $c$  and  $s$  in the symbol  $x_{b \cdot a}$  for the  $B$ -score associated with a fixed  $A$ -score or in the error components defined below: but we must distinguish between  $X_a = (x_a - M_a)$  and  $X_{a \cdot s} = (x_a - M_{a \cdot s})$  in virtue of the fact that the mean sample  $A$ -score  $M_{a \cdot cs} = M_{a \cdot s}$ , being that of the fixed- $A$  set, is not the same as the universe mean  $M_a = E_s(M_{a \cdot s})$ . With this convention, and with due regard to the fact that  $E_a$  in the prescribed order is a within-layer or within-stratum operation,

$$E_a(X_{a \cdot s}) = E_a(x_a - M_{a \cdot s}) = M_{a \cdot s} - M_{a \cdot s} = 0 \quad . \quad . \quad . \quad (ii)$$

$$E_a(X_a) = E_a(x_a - M_a) = M_{a \cdot s} - M_a \quad . \quad . \quad . \quad . \quad (iii)$$

In conceiving a universe stratified in this way, our end in view is to explore the consequences of a linear law relating the  $B$ -score means to the  $A$ -scores, i.e. linear regression of the  $B$ -score on the  $A$ -score. In doing so, we can take advantage of the assumptions inherent in the formulation of such a law, as stated in 17.00. That is to say, our  $B$ -scores have 2 additive *independent* components, the distribution of the *error* component about zero mean being *the same for every  $A$ -score* in any sub-universe specified by a fixed- $A$  set. When there is perfect linear regression in the stratified universe so conceived, we imply that

$$M_{b \cdot a} - M_b = k_{ba} \cdot X_a \quad \text{and} \quad M_{b \cdot a} = M_b + k_{ba} \cdot X_a \quad . \quad . \quad . \quad (iv)$$

If we express a physical law in linear form,  $M_{b \cdot a}$  in the above is the true value from which an observed  $B$ -score ( $x_{b \cdot a}$ ) deviates on account of instrumental error or imperfect control. Accordingly, we define our errors by the relations

$$\epsilon_{b \cdot a} = x_{b \cdot a} - M_{b \cdot a} = x_{b \cdot a} - M_b - k_{ba}(x_a - M_a) \quad . \quad . \quad . \quad (v)$$

$$x_{b \cdot a} = \epsilon_{b \cdot a} + M_{b \cdot a} = e_{b \cdot a} + M_b + k_{ba}(x_a - M_a) \quad . \quad . \quad . \quad (vi)$$

In conformity with this notation, we denote the mean *sample* error for a particular  $A$ -score as

$$M_{e \cdot acs} = E_{b \cdot a}(\epsilon_{b \cdot a}) = M_{b \cdot acs} - k_{ba} \cdot X_a - M_b;$$

$$M_{e \cdot as} = E_c(M_{e \cdot acs}) = M_{b \cdot as} - k_{ba} \cdot X_a - M_b = (M_{b \cdot a} - M_b) - k_{ba} \cdot X_a,$$

$$\therefore M_{e \cdot as} = 0 \quad \text{and} \quad M_{b \cdot acs} - M_{b \cdot as} = M_{e \cdot acs} \quad . \quad . \quad . \quad . \quad (vii)$$









Within the stratum  $V_{a.cs} = V_{a.s}$  is constant from layer to layer, so that

$$\frac{\text{Cov}(x_{a.cs}, x_{b.cs})}{V_{a.cs}} = k_{ba.cs} = \frac{\text{Cov}(x_{a.s}, x_{b.s})}{V_{a.s}},$$

$$\therefore E_c(k_{ba.cs}) = \frac{\text{Cov}(x_{a.s}, x_{b.s})}{V_{a.s}} = k_{ba.s} \quad \text{. . . . . (xxiv)}$$

(ii) The value of  $k_{ba.s}$  is the same in all strata, i.e. for the complete sampling distribution of any fixed- $A$  set. By (xii) above  $(M_{b.as} - M_{b.s}) = k_{ba.s} X_{a.s}$ . Whence from (xi) of 11.02:

$$\text{Cov}(x_{a.s}, x_{b.s}) = E_c.E_a(M_{b.as} - M_{b.s})(x_a - M_{a.s}) = k_{ba.s} E_c.E_a(x_a - M_{a.s})^2,$$

$$\therefore \text{Cov}(x_{a.s}, x_{b.s}) = k_{ba.s} V_{a.s}$$

$$\therefore k_{ba} = \frac{\text{Cov}(x_{a.s}, x_{b.s})}{V_{a.s}} = k_{ba.s} \quad \text{. . . . . (xxv)}$$

This result permits us to interpret (xxiv) in the form

$$E_c(k_{ba.cs}) = k_{ba} \quad \text{. . . . . (xxvi)}$$

This means that the mean of all samples of  $k_{ba.cs}$  in the fixed- $A$  set is the true regression coefficient. We express this by saying that the sample statistic which is an unbiased estimate of the true regression coefficient ( $k_{ba}$ ) of the parent universe is the ratio of the sample covariance to the sample variance of the  $A$ -score distribution.

If we assume *what is rarely true* (p. 714) of  $A$ -score sampling in experimental work, i.e. that our sample of  $A$ -scores is random, we can now derive the unbiased estimate of the true covariance. For random samples of  $p$  paired scores we may write

$$E_s.E_c(V_{a.cs}) = E_s(V_{a.s}) = \frac{p-1}{p} \sigma_a^2.$$

From (xxv) above

$$\text{Cov}(x_{a.s}, x_{b.s}) = k_{ba.s} V_{a.s} = k_{ba} V_{a.s},$$

$$\therefore E_s \text{Cov}(x_{a.s}, x_{b.s}) = k_{ba} E_s(V_{a.s}) = \frac{p-1}{p} \sigma_a^2 k_{ba},$$

$$\therefore E_s \text{Cov}(x_{a.s}, x_{b.s}) = \frac{(p-1) \text{Cov}(x_a, x_b)}{p},$$

$$\therefore E_s.E_c \text{Cov}(x_{a.cs}, x_{b.cs}) = \frac{(p-1) \text{Cov}(x_a, x_b)}{p}.$$

In sampling at random w.r.t. the  $A$ -score, the unbiased estimate of the covariance is therefore

$$\frac{p \text{Cov}(x_{a.cs}, x_{b.cs})}{(p-1)} = \sum_{j=1}^{j=p} \frac{(x_{aj} - M_{a.cs})(x_{bj} - M_{b.cs})}{(p-1)} \quad \text{. . . . . (xxvii)}$$

In what follows, we shall need to derive an estimate of the  $B$ -score variance of the stratum, which we may define by recourse to a 2-dimensional rearrangement of the cells in columns referable to a fixed  $A$ -score as

$$\begin{aligned} V_{b.s} &= M(V_{b.as}) + V(M_{b.as}) \\ &= M(V_{b.as}) + k_{ba.s}^2 V_{a.s} \\ &= M(V_{b.as}) + k_{ba}^2 V_{a.s}. \end{aligned}$$





To find the elastic constant we proceed as follows :

$$M_t = \frac{1}{4}(1 + 2 + 3 + 4) = 2.5;$$

$$M_s = \frac{1}{4}(0.4 + 1.1 + 1.4 + 2.1) = 1.25;$$

$$V_t = \frac{1}{4}(1 + 4 + 9 + 16) - (2.5)^2 = 1.25;$$

$$V_s = \frac{1}{4}(0.4^2 + 1.1^2 + 1.4^2 + 2.1^2) - (1.25)^2 = 0.3725;$$

$$\text{Cov}(s, t) = \frac{1(0.4) + 2(1.1) + 3(1.4) + 4(2.1)}{4} - (1.25)(2.5) = 0.675$$

$$k_{st} = \frac{0.675}{1.25} = 0.54.$$

In accordance with the notation we have used elsewhere (11.04 and 17.01), we may now formulate the regression equation which expresses the relation between the load ( $x_t$ ) and the *hypothetical regression score* ( $x_{r.t}$ ) which is our estimate of the corresponding stretch,

$$x_{r.t} - M_s = k_{st}(x_t - M_t) \quad \text{or} \quad x_{r.t} = k_{st} \cdot x_t + C \quad . \quad . \quad . \quad (ii)$$

The value of the constant definitive of the origin in (ii) is

$$C = M_s - k_{st} \cdot M_t.$$

In this case

$$C = 1.25 - 0.54(2.5) = 0.10.$$

The observed values and so-called *predicted* (i.e. assigned) values, i.e. values calculated in accordance with (i) and (ii) in agreement with what physicists call the least square method would thus be

<i>Observed</i>	0.4	1.1	1.4	2.1
<i>Predicted</i>	0.44	0.98	1.52	2.06

In this example the variance of the regression score distribution is

$$k_{st}^2 \cdot V_t = \frac{(0.675)^2}{1.25} = 0.3644.$$

As a fraction of the variance of the distribution of stretch scores this is

$$\frac{0.3644}{0.3725} \simeq 0.98.$$

We may thus say that a linear law here accounts for 98 per cent. of the variance of the stretch score distribution.

. . . . .

In the notation of 17.01 the regression equation expressing the so-called predicted value ( $x_{r.a}$ ) of a *B*-score for an *A*-score is

$$x_{r.a} = k_{ba.cs} \cdot x_a + C; \quad C = M_{b.cs} - k_{ba.cs} \cdot M_{a.s} \quad . \quad . \quad . \quad (iii)$$

It is important to distinguish the above from the exact relation connecting the observed *B*-score ( $x_b$ ) with the *A*-score. The appropriate equation which expresses it involves the true value of the regression coefficient and an error term, *viz.* :

$$x_b = k_{ba.s} \cdot x_a + C + x_e; \quad C = M_{b.s} - k_{ba.s} \cdot M_a \quad . \quad . \quad . \quad (iv)$$

For speedy computation of  $k_{ba.cs}$  we may proceed as follows for  $p$  paired values :

$$\sum_{j=1}^{j=p} x_{a.j} = s_a = p \cdot M_a; \quad \sum_{j=1}^{j=p} x_{a.j}^2 = s_{aa} \quad . \quad . \quad . \quad . \quad (v)$$

$$\sum_{j=1}^{j=p} x_{b.j} = s_b = p \cdot M_b; \quad \sum_{j=1}^{j=p} x_{b.j}^2 = s_{bb} \quad . \quad . \quad . \quad . \quad (vi)$$

$$\sum_{j=1}^{j=p} x_{a.j} \cdot x_{b.j} = s_{ab} \quad . \quad . \quad . \quad . \quad (vii)$$

In this notation

$$\text{Cov}(x_a, x_b) = \frac{s_{ab}}{p} - \frac{s_a s_b}{p^2};$$

$$V_a = \frac{s_{aa}}{p} - \frac{s_a^2}{p^2}.$$

Whence in (iii) above

$$k_{ba.cs} = \frac{p \cdot s_{ab} - s_a \cdot s_b}{p \cdot s_{aa} - s_a^2} \quad . \quad . \quad . \quad . \quad (viii)$$

$$C = \frac{1}{p} (s_b - k_{ba.cs} \cdot s_a) \quad . \quad . \quad . \quad . \quad (ix)$$

The computation sheet will thus require five columns :

	$x_a$	$x_b$	$x_a^2$	$x_b^2$	$x_a \cdot x_b$
	...	...	...	...	...
Totals	$s_a$	$s_b$	$s_{aa}$	$s_{bb}$	$s_{ab}$

### 17.03 UNBIASED ESTIMATES OF REGRESSION PARAMETERS

Having established the conclusion that the ratio of the sample covariance of the bivariate distribution to the sample variance of the  $A$ -score distribution is an unbiased estimate of the slope constant (regression coefficient) if the mean  $B$ -score of the *u.s.d.* is a linear function of the  $A$ -score, we have as yet no criterion of the sampling error of the regression coefficient nor of the sampling error of a  $B$ -score estimate based on its sample value ; and we have still to justify our confidence that the true law of the universe is linear when a sample of paired scores furnishes us with the only available precise information about its structure.

Before attempting to answer the questions implicit in the last two sentences, we may with advantage retrace our steps to the significance test for the correlation ratio in 16.08. This test purports to answer the question : have we good reason for believing that there is a law asserting the dependence of the  $B$ -score on the  $A$ -score ? When they take the next step by asking whether the form of the law is linear, exponents of Fisher's test procedures follow the same path inasmuch as they seek to define what different estimates of the variance of the  $B$ -score distribution or that of its error components must be consistent, if the null hypothesis is correct.

Since the significance tests which we shall now examine lean heavily on those developed in Chapter 16 in connexion with the *Analysis of Variance*, it is helpful to view the sample structure



as a score-grid on all fours with the customary schema of 2 columns and  $p$  rows for  $p$  paired score values. Accordingly, we recall the illustration (p. 431) at the end of 11.01. Below is a symmetrical frequency grid exhibiting 16 paired scores in the range 1 to 4. We shall also lay it out as a score-grid (Table 1).

		<i>A</i> -score				Total	Sum	Mean
		1	2	3	4			
<i>B</i> -score	1	1	1	.	.	2	3	$\frac{9}{6}$
	2	1	3	2	.	6	13	$\frac{13}{6}$
	3	.	2	3	1	6	17	$\frac{17}{6}$
	4	.	.	1	1	2	7	$\frac{21}{6}$
Total		2	6	6	2	16	40	$\frac{5}{2}$
Sum		3	13	17	7	40	...	..
Mean		$\frac{9}{6}$	$\frac{13}{6}$	$\frac{17}{6}$	$\frac{21}{6}$	$\frac{5}{2}$	...	..

The alternative lay-out calls for a modification of our notation exhibited in the accompanying score-grid on the assumption that our concern is with the regression of the *B*-score on the *A*-score. We then need to distinguish one *A*-score ( $a_i$ ) from another in virtue of its numerical value only, indicated by the rank subscript  $i$ , but we need also to distinguish *B*-scores ( $b_{j.i}$ ) associated with the same value of the *A*-score and label them with a double subscript accordingly.

TABLE 1

<i>A</i> -score ( $a_i$ )	<i>B</i> -score ( $b_{j \cdot i}$ )	For fixed value of <i>A</i> -score	
		Total of <i>B</i> -scores ( $T_{b \cdot i}$ )	No. of <i>B</i> -scores ( $p_i$ )
$a_1 = 1$	$b_{1 \cdot 1} = 1$	$\sum_{j=1}^{j=2} b_{j \cdot 1} = T_{b \cdot 1} = 3$	$p_1 = 2$
$a_1 = 1$	$b_{2 \cdot 1} = 2$		
$a_2 = 2$	$b_{1 \cdot 2} = 1$	$\sum_{j=1}^{j=6} b_{j \cdot 2} = T_{b \cdot 2} = 13$	$p_2 = 6$
$a_2 = 2$	$b_{2 \cdot 2} = 2$		
$a_2 = 2$	$b_{3 \cdot 2} = 2$		
$a_2 = 2$	$b_{4 \cdot 2} = 2$		
$a_2 = 2$	$b_{5 \cdot 2} = 3$		
$a_2 = 2$	$b_{6 \cdot 2} = 3$		
$a_3 = 3$	$b_{1 \cdot 3} = 2$	$\sum_{j=1}^{j=6} b_{j \cdot 3} = T_{b \cdot 3} = 17$	$p_3 = 6$
$a_3 = 3$	$b_{2 \cdot 3} = 2$		
$a_3 = 3$	$b_{3 \cdot 3} = 3$		
$a_3 = 3$	$b_{4 \cdot 3} = 3$		
$a_3 = 3$	$b_{5 \cdot 3} = 3$		
$a_3 = 3$	$b_{6 \cdot 3} = 4$		
$a_4 = 4$	$b_{1 \cdot 4} = 3$	$\sum_{j=1}^{j=2} b_{j \cdot 4} = T_{b \cdot 4} = 7$	$p_4 = 2$
$a_4 = 4$	$b_{2 \cdot 4} = 4$		
$T_a = \sum_{i=1}^{i=4} p_i a_i = 40$	$\sum_{i=1}^{i=4} \sum_{j=1}^{j=p_i} b_{j \cdot i} = \sum_{i=1}^{i=4} T_{b \cdot i} = 40$		$p = \sum_{i=1}^{i=4} p_i = 16$
$M_a = \frac{T_a}{16}$	$M_b = \frac{T_b}{16}$		.....

In the accompanying schema we introduce no conventions to distinguish sample stratum (fixed-*A* set) or universe (whole 3-dimensional grid) parameters. Where necessary we can do this as in 17.01, e.g. for mean *B*-scores associated with the fixed *A*-score  $a_i$  we may write  $M_{b \cdot i c s}$ ,  $M_{b \cdot i s}$ ,  $M_{b \cdot i}$ . For the present, this need not concern us. It will suffice to write

$$M_a = \frac{1}{p} \sum_{i=1}^{i=c} p_i a_i; \quad M_{b \cdot i} = \frac{1}{p_i} \sum_{j=1}^{j=p_i} b_{j \cdot i};$$

$$\frac{1}{p} \sum_{i=1}^{i=c} \sum_{j=1}^{j=p_i} b_{j \cdot i} = M_b = \frac{1}{p} \sum_{i=1}^{i=c} p_i \cdot M_{b \cdot i};$$



$$\begin{aligned}
V_{b \cdot i} &= \frac{1}{p_i} \sum_{j=1}^{j=p_i} (b_{j \cdot i} - M_{b \cdot i})^2 = \frac{1}{p_i} \sum_{j=1}^{j=p_i} (b_{j \cdot i} - M_b)^2 - (M_{b \cdot i} - M_b)^2; \\
M(V_{b \cdot i}) &= \frac{1}{p} \sum_{i=1}^{i=c} p_i V_{b \cdot i} = \frac{1}{p} \sum_{i=1}^{i=c} \sum_{j=1}^{j=p_i} (b_{j \cdot i} - M_b)^2 - \frac{1}{p} \sum_{i=1}^{i=c} p_i (M_{b \cdot i} - M_b)^2 \\
&= V_b - V(M_{b \cdot i}); \\
Cov(a_i, b_{j \cdot i}) &= \frac{1}{p} \sum_{i=1}^{i=c} p_i (a_i - M_a)(M_{b \cdot i} - M_b).
\end{aligned}$$

When we employ this notation henceforth, we shall write  $(a_i - M_a) = A_i$ , so that

$$\sum_{i=1}^{i=c} p_i \cdot A_i = 0 \quad \text{and} \quad \sum_{i=1}^{i=c} p_i \cdot A_i^2 = V_a \quad . \quad . \quad . \quad . \quad (i)$$

$$Cov(a_i, b_{j \cdot i}) = \frac{1}{p} \sum_{i=1}^{i=c} p_i \cdot A_i (M_{b \cdot i} - M_b) \quad . \quad . \quad . \quad . \quad (ii)$$

In what follows, we shall need to recall an important property of covariances, *viz.* that

$$\begin{aligned}
Cov(a_i, b_{j \cdot i}) &= \frac{1}{p} \sum_{i=1}^{i=c} p_i \cdot A_i \cdot M_{b \cdot i} - \frac{M_b}{p} \sum_{i=1}^{i=c} p_i \cdot A_i \\
&= \frac{1}{p} \sum_{i=1}^{i=c} p_i \cdot A_i \cdot M_{b \cdot i} \quad . \quad . \quad . \quad . \quad (iii)
\end{aligned}$$

Within the fixed- $A$  set,  $A_{i \cdot s}$  has the meaning  $(a_i - M_{a \cdot s})$ . We may therefore write

$$\begin{aligned}
k_{ba \cdot cs} &= \sum_{i=1}^{i=c} \frac{p_i \cdot A_{i \cdot s} \cdot M_{b \cdot ics}}{p \cdot V_{a \cdot s}} \quad \text{and} \quad k_{ba \cdot s} = \sum_{i=1}^{i=c} \frac{p_i \cdot A_{i \cdot s} \cdot M_{b \cdot is}}{p \cdot V_{a \cdot s}} = k_{ba}, \\
\therefore (k_{ba \cdot cs} - k_{ba})p \cdot V_{a \cdot s} &= \sum_{i=1}^{i=c} p_i \cdot A_{i \cdot s} (M_{b \cdot ics} - M_{b \cdot is}) \quad . \quad . \quad (iv)
\end{aligned}$$

Whence from (vii) of 17.01

$$(k_{ba \cdot cs} - k_{ba})p \cdot V_{a \cdot s} = \sum_{i=1}^{i=c} p_i \cdot A_{i \cdot s} \cdot M_{e \cdot ics} \quad . \quad . \quad . \quad (v)$$

So long as we restrict our attention to the fixed- $A$  set, we view the sampling process through the spectacles of Churchill Eisenhart's Model I in 13.04. That is to say, we conceive our sample as one of an endless repetition of experiments *done in the same way*. In the idiom of 17.01, our bivariate universe is thus a stratum with all possible values of the  $B$ -score distribution for a *fixed- $A$  set*; and the student must therefore lay aside preoccupations suggested by any previous acquaintance with the so-called bivariate normal universe. With more or less propriety, we may assume an approximately normal distribution of the  $B$ -scores for any fixed value of  $A$ ; but the distribution of the  $A$ -scores will rarely be approximately normal—and indeed calls for no explicit specification—in the infinite succession of similar experiments among which our actual sample constitutes a single act of repetition. Throughout what follows, as in later sections of this chapter, we therefore assume consistently that the relevant framework of repetition is a fixed- $A$  set which may have any distribution involving at least 3 different  $A$ -score values. This assumption is unnecessary only when  $k_{ba} = 0 = r_{ba}$  in the universe as a whole. In effect, we then sample





Whence from (vii) above :

$$\sigma_{k.s}^2 = \frac{\sigma_e^2}{p \cdot V_{a.s}} = \frac{\sigma_e^2}{\sum_{i=1}^c p_i \cdot A_i^2} \quad \dots \quad (x)$$

If we now make the assumption that the error distribution is normal, whence that of the mean of a  $p_i$ -fold sample is a normal variate, the principle of the fixed- $A$  set implies that  $(k_{ba.cs} - k_{ba})$  as defined by (viii) is a sum of independent normal variates and is itself therefore a normal variate with zero mean, since  $k_{ba}$  is the mean value of  $k_{ba.cs}$ . Thus we may define a normal square standard score (Chi-Square variate of 1 d.f.) by the ratio

$$\frac{(k_{ba.cs} - k_{ba})^2}{\sigma_{k.s}^2} = \frac{(k_{ba.cs} - k_{ba})^2 p \cdot V_{a.s}}{\sigma_e^2} \quad \dots \quad (xi)$$

The error variance ( $\sigma_e^2$ ) is, of course, the expected mean square deviation of the  $B$ -score from its true mean value defined by the relation  $M_{b.a} = k_{ba}X_a + M_b$ . All our sample tells us is the deviation of the  $B$ -score from the hypothetical regression score

$$x_{r.a} = k_{ba.cs}X_{a.s} + M_{b.cs}$$

To get an unbiased estimate of  $\sigma_e^2$ , we therefore examine the implications of the tautology of (xxi) in 11.04. In the notation of 17.01, this is

$$E_a E_{b.a} (x_{b.a} - x_{r.acs})^2 = V_{b.cs} - k_{ba.cs}^2 V_{a.s} = (1 - r_{ba.cs}^2) V_{b.cs} \quad \dots \quad (xii)$$

Within the fixed- $A$  set,  $V_{a.s}$  is a constant. Thus the expected value of the mean square deviation of the  $B$ -score from the sample regression line within the fixed- $A$  set is

$$E_c(V_{b.cs}) - V_{a.s} \cdot E_c(k_{ba.cs})^2.$$

By (xxix) of 17.01, this is

$$\frac{p-1}{p} \sigma_e^2 + k_{ba}^2 V_{a.s} - V_{a.s} \cdot E_c(k_{ba.cs})^2 \quad \dots \quad (xiii)$$

Now by definition,

$$E_c(k_{ba.cs} - k_{ba})^2 = \sigma_{k.s}^2 = E_c(k_{ba.cs})^2 - k_{ba}^2, \\ \therefore E_c(k_{ba.cs})^2 V_{a.s} = \sigma_{k.s}^2 V_{a.s} + k_{ba}^2 V_{a.s}.$$

Whence (xiii) becomes

$$\frac{p-1}{p} \sigma_e^2 - \sigma_{k.s}^2 V_{a.s}.$$

Also from (x) this is

$$\frac{p-1}{p} \sigma_e^2 - \frac{\sigma_e^2}{p} = \frac{p-2}{p} \sigma_e^2 \quad \dots \quad (xiv)$$

We may thus define a statistic whose expected value is  $\sigma_e^2$  by the relations

$$s_e^2 = \frac{p}{p-2} E_a \cdot E_{b.a} (x_{b.a} - x_{r.acs})^2 \text{ and } E_s(s_e^2) = \sigma_e^2 \quad \dots \quad (xv)$$

Alternatively, we may write

$$s_e^2 = \frac{(1 - r_{ab.cs}^2) p \cdot V_{b.cs}}{p-2} \quad \dots \quad (xvi)$$





## QUADRATIC REGRESSION IN A BIVARIATE UNIVERSE

(Sampling with replacement)

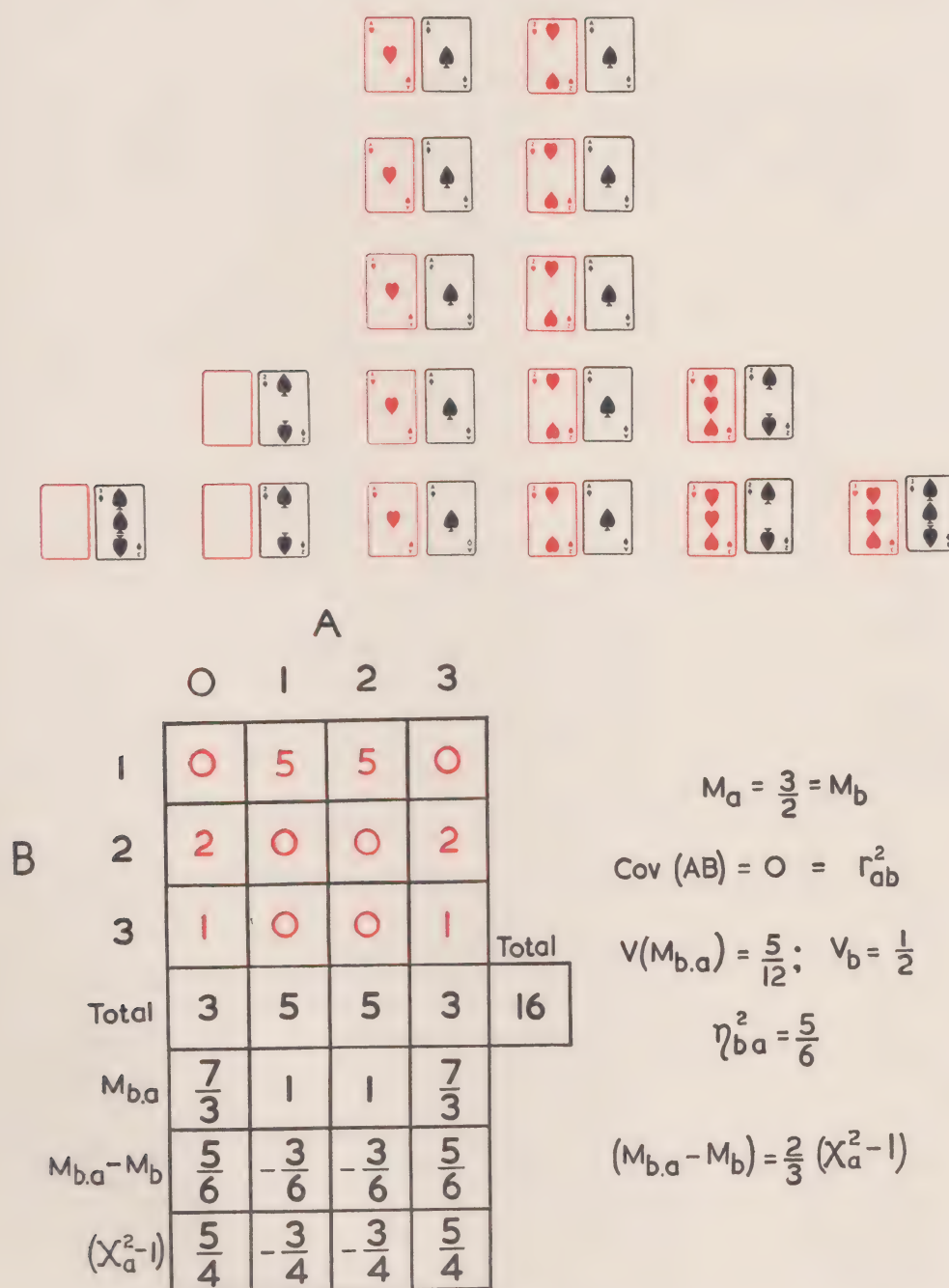


FIG. 122.

The approximate test defined by (xix) involves the ratio of two complementary fractions of the variance of the  $B$ -score distribution, *viz.*  $r_{ab}^2 \cdot V_b$  and  $(1 - r_{ab}^2)V_b$ . Significance tests prescribed by Fisher's school involve the derivation of certain relations relevant to other partitions of the variance of the  $B$ -score distribution. We have already examined one such in 16.08, where we exhibited a significance test for the correlation ratio ( $\eta_{ba}^2$ ). If the result of the latter test can show that the correlation ratio does not differ significantly from zero, there is, of course, no need to ask whether the product-moment index significantly exceeds zero. We have seen

(footnote to 11.04) that  $\eta_{ba}^2$  must be greater than  $r_{ab}^2$  or equal to it, being equal if regression is linear, including the trivial case  $r_{ab}^2 = 0$ . If the true value of  $\eta_{ba}^2$  is greater than that of  $r_{ba}^2$ , there must be a *non-linear* relation between the two sets of scores. Fig. 122 shows that a high value of  $\eta_{ba}^2$  is indeed consistent with zero covariance when the *B*-score means are a quadratic function of the *A*-scores.

To say that  $\eta_{ba}^2$  significantly exceeds zero and that  $r_{ab}^2$  does not, thus points to the existence of a non-linear dependence of the *B*-score on the *A*-score. This gives us the clue to a way of testing whether regression is in fact linear. We require to know the distribution of a sample statistic which measures the excess of  $\eta_{ba}^2$  over  $r_{ab}^2$ . To this end we shall first recall (Table 2)

TABLE 2

*Tautologies of Regression*

Source	Parameter
1. Sampling Error	$M(V_{b \cdot acs}) = (1 - \eta_{ba \cdot cs}) V_{b \cdot cs}$ $E_a E_{b \cdot a} (x_{b \cdot a} - M_{b \cdot acs})^2 = \frac{1}{p} \sum_{i=1}^{i=c} \sum_{j=1}^{j=p_i} (b_{j \cdot i} - M_{b \cdot iccs})^2$
2. Non-linear Regression	$V(M_{b \cdot acs}) - k_{ba \cdot cs}^2 \cdot V_{a \cdot s} = (\eta_{ba \cdot cs}^2 - r_{ba \cdot cs}^2) V_{b \cdot cs}$ $E_a (M_{b \cdot acs} - x_{r \cdot acs})^2 = \frac{1}{p} \sum_{i=1}^{i=c} p_i (M_{b \cdot iccs} - x_{r \cdot iccs})^2$
3. Linear Regression	$k_{ba \cdot cs}^2 \cdot V_{a \cdot s} = r_{ab \cdot cs}^2 \cdot V_{b \cdot cs}$ $E_a (x_{r \cdot acs} - M_{b \cdot cs})^2 = \frac{1}{p} \sum_{i=1}^{i=c} p_i (x_{r \cdot iccs} - M_{b \cdot cs})^2$
4. Influence of <i>A</i> on <i>B</i>	$V(M_{b \cdot acs}) = \eta_{ba \cdot cs}^2 \cdot V_{b \cdot cs}$ $E_a (M_{b \cdot acs} - M_{b \cdot cs})^2 = \frac{1}{p} \sum_{i=1}^{i=c} p_i (M_{b \cdot iccs} - M_{b \cdot cs})^2$
5. Error	$(1 - r_{ba \cdot cs}) V_{b \cdot cs}$ $E_a E_{b \cdot a} (x_{b \cdot a} - x_{r \cdot acs})^2 = \frac{1}{p} \sum_{i=1}^{i=c} \sum_{j=1}^{j=p_i} (b_{j \cdot i} - M_{b \cdot cs} - k_{ba \cdot cs} A_i)^2$
6. Total of 1, 2 and 3, of 1 and 4 and of 3 and 5	$V_{b \cdot cs}$ $E_a E_{b \cdot a} (x_{b \cdot a} - M_{b \cdot cs})^2 = \frac{1}{p} \sum_{i=1}^{i=c} \sum_{j=1}^{j=p_i} (b_{j \cdot i} - M_{b \cdot cs})^2$

as in the accompanying table certain necessary relations between *B*-scores, hypothetical linear regression scores and *B*-score means already exhibited as tautologies of the grid in 11.04. The column headed *source* in Table 2 anticipates the statistical interpretation we shall later impose on each parameter defined alternatively in the notation of this section and in that of 17.01.

To make explicit what information about the universe each sample parameter of Table 1 supplies in terms of the source of variation, we must explore what is its *expected* value. The



hypothesis stated in 17.00 is that each observed  $B$ -score has one component, the true value, which is a linear function of the  $A$ -score, and an independent (*error*) component ( $e_{j,i}$ ) with zero mean. This is on all fours with the consequential relation between the scores of the player and umpire in the model set-up of 12.01, where we postulate a relation of the type  $x_a = k \cdot x_u + x_{a.o}$ . We may translate our hypothesis into the language of 13.04, if we conceive the true value as a *regression-factor* in the 2 column grid of paired scores, i.e.

$$F_i = k_{ba} \cdot a_i + C \quad \text{and} \quad b_{j..i} = F_j + e_{j..i}. \quad . \quad . \quad . \quad (\text{xxiii})$$

The first 6 items of the second column score-grid for the 16 pair set exhibited in Table 1 then take the form

$$\begin{array}{ll} a_1 & b_{1.1} = F_1 + e_{1.1} \\ a_1 & b_{2.1} = F_1 + e_{2.1} \\ a_2 & b_{1.2} = F_2 + e_{1.2} \\ a_2 & b_{2.2} = F_2 + e_{2.2} \\ a_2 & b_{3.2} = F_2 + e_{3.2} \\ a_2 & b_{4.2} = F_2 + e_{4.2} \end{array}$$

The mean error is zero for any value of the  $A$ -score in the complete sample distribution of the fixed- $A$  set. For a fixed value of the  $A$ -score within the sample and within the sample distribution of the fixed- $A$  set, we may therefore write

$$M_{b,i,c_s} = F_i + M_{e,i,c_s} \quad \text{and} \quad M_{b,i,s} = F_i + M_{e,i,s} = F_i \quad . \quad . \quad (\text{xxiv})$$

If the mean stratum value of  $F_i$  is  $M_{f.s.}$ , we may therefore write :

$$M_{b.cs} = M_{f.s} + M_{e.cs} \quad \text{and} \quad M_{b.s} = M_{f.s};$$

$$M_{b.is} - M_{b.cs} = (F_i - M_{f.s}) + (M_{e.is} - M_{e.cs}) \quad \text{and} \quad M_{b.is} - M_{b.s} = (F_i - M_{f.s}) \quad (\text{xxv})$$

In the last expressions,

$$F_i - M_{f.s} = (k_{ba} \cdot a_i + C) - (k_{ba} \cdot M_{a.s} + C) = k_{ba}(a_i - M_{a.s}) = k_{ba} \cdot A_{i.s}.$$

Thus (xxiv) expresses the relations defined by (xii) in 17.01 in terms of the factor concept of 13.04. In accordance with our treatment of the Model 1 balance sheet of variance, we shall therefore write the variance of the  $A$ - $B$  factor distributions :

[illegible]

The sample value of the  $B$ -score variance for a fixed  $A$ -score depends only on the error component, since the regression-factor is then fixed, i.e.  $V_{b \cdot is} = V_{e \cdot is}$ . If there are  $p_i$   $B$ -scores associated with one and the same value ( $a_i$ ) of the  $A$ -score, we may write the expected value of  $V_{b \cdot is}$  in the notation of 13.04 as

$$E_c(V_{b.a.c.s.}) = \frac{p_i - 1}{p_i} V_e.$$

In the same notation the expected value of the *mean* variance of the *B*-score distribution is

$$\begin{aligned} E_c \cdot M(V_{b \cdot acs}) &= E_c \cdot E_a(V_{b \cdot acs}) = E_a \cdot E_c(V_{b \cdot acs}), \\ \therefore E_c \cdot M(V_{b \cdot acs}) &= \sigma_e^2 E_a \left( \frac{p_i - 1}{p_i} \right). \end{aligned}$$





TABLE 3  
*Estimates of Regression*

Source of Variation	Sample Parameter	Expected Value	Adjusted Sample Statistic	Expected Value
1. Error	$(1 - \eta_{ba \cdot cs}^2)V_{b \cdot cs}$	$\frac{(p-c)}{p} \sigma_e^2$	$s_u^2 = \frac{(1 - \eta_{ba \cdot cs}^2)p \cdot V_{b \cdot cs}}{p - c}$	$\sigma_e^2$
2. Non-linear Regression	$(\eta_{ba \cdot cs}^2 - r_{ba \cdot cs}^2)V_{b \cdot cs}$	$\frac{c-2}{p} \sigma_e^2 + (\sigma_r^2 - k_{ba} \cdot V_{a \cdot s})$	$s_c^2 = \frac{(\eta_{ba \cdot cs}^2 - r_{ba \cdot cs}^2)p \cdot V_{b \cdot cs}}{c - 2}$	$> \sigma_e^2$ (if regression is not linear) $\sigma_e^2$ (if regression is linear)
3. Linear Regression	$r_{ba \cdot cs}^2 V_{b \cdot cs}$	$\frac{\sigma_e^2}{p} + \sigma_r^2$ if regression is linear	$s_r^2 = r_{ba \cdot cs}^2 \cdot p \cdot V_{b \cdot cs}$	$\sigma_e^2 + p\sigma_r^2$ (= $\sigma_e^2$ if $k_{ba} = 0$ )
4. Unspecified Dependence	$\eta_{ba \cdot cs}^2 \cdot V_{b \cdot cs}$	$\frac{(c-1)}{p} \sigma_e^2 + \sigma_r^2$	$s_m^2 = \frac{\eta_{ba \cdot cs}^2 p \cdot V_{b \cdot cs}}{c - 1}$	$\sigma_e^2 + \frac{p}{c-1} \sigma_r^2$ (= $\sigma_e^2$ if $\sigma_r^2 = 0$ )
5. Linear Residual	$(1 - r_{ba \cdot cs}^2)V_{b \cdot cs}$	$\frac{p-2}{p} \sigma_e^2$	$s_e^2 = \frac{(1 - r_{ba \cdot cs}^2)p \cdot V_{b \cdot cs}}{p - 2}$	$\sigma_e^2$ if regression is linear
6. Total Variance (Sum of 1, 2 and 3, of 1 and 4 and of 3 and 5)	$V_{b \cdot cs}$	$\frac{p-1}{p} \sigma_e^2 + \sigma_r^2$	$s_t^2 = \frac{p \cdot V_{b \cdot cs}}{p - 1}$	$\sigma_e^2 + \frac{p}{p-1} \sigma_r^2$

## 17.04 SIGNIFICANCE OF ESTIMATED REGRESSION PARAMETERS

Table 3 of 17.03 exhibits various sample statistics, one ( $s_u^2$ ) of which is necessarily an unbiased estimate of residual variance ( $\sigma_e^2$ ). The expected value of others will necessarily exceed the latter unless one or other of certain conditions specified in the column at the extreme right holds good. The expected value of the ratio of any one of them to  $s_u^2$  will then exceed the expected value of the ratio of consistent estimates of variance. We have explored the distribution of such a ratio on the assumption that : (a) the score distribution in the parent universe is normal ; (b) the two variances are statistically independent. Before we can employ the sample statistics of Table 3 as a basis for testing whether one or other prescribed condition does in fact hold good, it will therefore be necessary to examine which we can pair off as independent statistics.

The student who recalls the test for the significance of a correlation ratio in 16.08 will indeed recognise in Table 3 of 17.03 two independent estimates whose consistency is a criterion of the dependence of the  $B$ -score on the  $A$ -score, *viz.* those here denoted  $s_m^2$  and  $s_u^2$ . We may denote their ratio as

$$F_{mu} = \frac{\eta_{ba.cs}^2}{1 - \eta_{ba.cs}^2} \cdot \frac{p - c}{c - 1}.$$

If (and only if)  $\sigma_e^2 = \sigma_b^2$ , as when  $\sigma_r^2 = 0$ , this is the ratio of 2 independent Chi-Square variates of  $(c - 1)$  and  $(p - c)$  degrees of freedom respectively, being therefore a Type VI variate. Otherwise, the expected value of  $\eta_{ba.cs}^2$  will be greater than that of  $(1 - \eta_{ba.cs}^2)$ . We must therefore regard an uncommonly *high* value of  $F_{mu}$  as : (i) a rarity if we are content to accept the null hypothesis that there is no causal nexus involved in the pair score distribution ; (ii) alternatively, as ground for dismissing the validity of the null hypothesis.

Scrutiny of Table 3 invites examination of the properties of two other ratios, one as a criterion of zero covariance ( $k_{ba} = 0$ ), the other of linear regression, *viz.* :

$$F_{re} = \frac{(p - 2) \cdot r_{ba.cs}^2}{(1 - r_{ba.cs}^2)} \text{ and } F_{cu} = \frac{(\eta_{ba.cs}^2 - r_{ba.cs}^2)}{(1 - \eta_{ba.cs}^2)} \cdot \frac{(p - c)}{(c - 2)} \quad . \quad . \quad . \quad (i)$$

In connexion with any use we may subsequently make of  $F_{mu}$  above and of either ratio in (xxiv), one property of the denominator calls for comment. If we have only a single  $B$ -score for each different value of the  $A$ -score, the sample variance of each column in the frequency grid is zero and  $M(V_{b.acs}) = 0$ . Such a unique relation between  $A$ -score and  $B$ -score values also signifies  $p = c$ , so that  $(p - c) = 0$ . Thus the statistic  $s_u^2$  in Table 3 is indeterminate, being the ratio of two zeros. A desideratum of the three significance tests based on  $s_u^2$  and discussed below is therefore that there are several non-identical  $B$ -scores for at least one value of the  $A$ -score.

To get into focus the problems we are now ready to tackle we may with profit recall the assumptions of the test based on  $F_{mu}$  above as stated in 16.08. There the null hypothesis was that each set of  $B$ -scores corresponding to a particular  $A$ -score is a sample from the same universe as the set of  $B$ -scores associated with any other value of the  $A$ -score. If we look on each column of the score grid of 16.08 as a sample from a sub-universe of  $B$ -scores, our postulate is therefore that each such sub-universe is identical with any other. Formally this means that  $F_i = 0$  in our equation of  $B$ -score build-up, so that  $b_{j.i} = \epsilon_{j.i} + C$ . The constant  $C$  which is common to all  $B$ -scores regardless of the meaning attached to  $i$  then merely signifies that the  $B$ -score mean, unlike that of the  $e$ -score component, is not (necessarily) zero. If our null hypothesis includes the assumption that regression is linear but excludes the assumption that the  $B$ -scores and  $A$ -scores are independent, our model universe is no longer *Bernoullian* (i.e. homogeneous).



Each column of the grid of 16.08 is then a unique sub-universe. The entire universe of  $B$ -scores is then a stratified universe.

To define the commonly prescribed  $F$ -test for linearity, and a so-called *exact* test of significance based on (xix) of 17.03 we have to be clear about three things :

- (a) any such test assumes a normal distribution of the error component of the  $B$ -scores and is more or less exact only in so far as this postulate is more or less correct ;
- (b) whereas the universe of  $e$ -scores is homogeneous by hypothesis, sampling in the universe of  $B$ -scores is *stratified* unless the  $B$ -scores and  $A$ -scores are independent, in which case  $\sigma_b^2 = \sigma_e^2$ , and the issue of linearity does not arise ;
- (c) a Chi-Square variate is the sum of square standard scores of unit variance and as such involves the true and *unknown* value of the appropriate universe parameter which will disappear in the variance ratio only if it appears in both the numerator and denominator of the latter.

We cannot legitimately employ the customary formula ( $\sigma_m^2 = \sigma^2 \div n$ ) for the variance of the mean unless we are sampling in a homogeneous universe, in this case the universe of  $e$ -scores ; and the denominator ( $1 - \eta_{ba.cs}^2$ ) of  $F_{cu}$  in (i) is, for reasons we shall state more explicitly at a later stage, a parameter of the  $e$ -score distribution alone. Consequently, (c) implies that any sample parameters of an  $F$ -ratio we invoke within the framework of the assumption that  $B$ -scores and  $A$ -scores are not necessarily independent must be expressible in terms of the  $e$ -score distribution alone.

Our next task must therefore be to define in what circumstances it is possible to express the sample statistics ( $\eta_{ba.cs}^2 - r_{ba.cs}^2$ ) in the numerator of  $F_{cu}$  as defined by (i) above and  $s_e^2$  as defined by (xvi) of 17.03 in terms which involve only error components and constants. We first recall the tautologies (Table 2) :

$$(\eta_{ba.cs}^2 - r_{ba.cs}^2)V_{b.cs} = E_a(M_{b.acs} - x_{r.acs})^2 = E_a(M_{b.acs} - M_{b.cs} - k_{ba.cs} \cdot X_{a.s})^2 \quad (ii)$$

$$(1 - r_{ba.cs}^2)V_{b.cs} = E_a \cdot E_{b.a}(x_{b.a} - x_{r.acs})^2 = E_a(x_{b.a} - M_{b.cs} - k_{ba.cs} \cdot X_{a.s})^2 \quad (iii)$$

We shall also need to make use of (vi) and (viii) 17.03, *viz.* :

$$p \cdot V_{a.s}(k_{ba.cs} - k_{ba}) = \sum_{i=1}^{i=c} p_i \cdot A_i \cdot M_{e.ics} = p \cdot E_a(X_{a.s} \cdot M_{e.acs}) \quad (iv)$$

If regression is linear,  $M_{b.as} = k_{ba} \cdot X_{a.s} + M_{b.s}$ . Also, in any case

$$(M_{b.acs} - M_{b.as}) = M_{e.acs} \quad \text{and} \quad (M_{b.cs} - M_{b.s}) = M_{e.cs}$$

by (vii) and (xi) of 17.01, whence we can write (ii) in the form

$$\begin{aligned} (\eta_{ba.cs}^2 - r_{ba.cs}^2)V_{b.cs} &= E_a[(M_{b.acs} - M_{b.as}) - (M_{b.cs} - M_{b.s}) - (k_{ba.cs} - k_{ba})X_{a.s}]^2 \\ &= E_a[(M_{e.acs} - M_{e.cs}) - (k_{ba.cs} - k_{ba})X_{a.s}]^2 \\ &= E_a(M_{e.acs}^2) + (M_{e.cs}^2) + (k_{ba.cs} - k_{ba})^2 V_{a.s} \\ &\quad - 2M_{e.cs} \cdot E_a(M_{e.acs}) \\ &\quad + 2M_{e.cs}(k_{ba.cs} - k_{ba}) \cdot E_a(X_{a.s}) \\ &\quad - 2(k_{ba.cs} - k_{ba}) \cdot E_a(X_{a.s} \cdot M_{e.acs}). \end{aligned}$$

In this expression  $E_a(M_{e.acs}) = M_{e.cs}$  and  $E_a(X_{a.s}) = 0$ , whence from (iv)

$$(\eta_{ba.cs}^2 - r_{ba.cs}^2)V_{b.cs} = E_a(M_{e.acs}^2) - M_{e.cs}^2 - (k_{ba.cs} - k_{ba})^2 V_{a.s} \quad (v)$$

We may write this alternatively as

$$\frac{(\eta_{ba.cs}^2 - r_{ba.cs}^2)p \cdot V_{b.cs}}{\sigma_e^2} = \frac{\sum_{i=1}^{i=c} p_i \cdot M_{e.ic}^2}{\sigma_e^2} - \frac{p \cdot M_{e.cs}^2}{\sigma_e^2} - \frac{(k_{ba.cs} - k_{ba})^2 p \cdot V_{a.s}}{\sigma_e^2}.$$

Since our universe of  $e$ -score components is homogeneous, we may write the variances of the mean  $e$ -score for a sample of  $p_i$  associated with a fixed value of the  $A$ -score, and for a  $p$ -fold sample associated with all the  $A$ -scores of the fixed set, respectively as

$$\sigma_{m.i}^2 = \frac{\sigma_e^2}{p_i} \quad \text{and} \quad \sigma_m^2 = \frac{\sigma_e^2}{p}.$$

Whence the foregoing expression becomes

$$\frac{(\eta_{ba.cs}^2 - r_{ba.cs}^2)p \cdot V_{b.cs}}{\sigma_e^2} = \sum_{i=1}^{i=c} \frac{M_{e.ic}^2}{\sigma_{m.i}^2} - \frac{M_{e.cs}^2}{\sigma_m^2} - \frac{(k_{ba.cs} - k_{ba})p \cdot V_{a.s}}{\sigma_e^2}. \quad (\text{vi})$$

Since the true mean of the errors for a fixed value of the  $A$ -score or for any fixed- $A$  set as a whole is zero, the first term on the right is the true variance of the  $e$ -score means for a fixed  $A$ -score being therefore a Chi-Square variate of  $c$  degrees of freedom. The second is a Chi-Square variate of 1 d.f. and the third is, being as already shown a square standard normal score, also a Chi-Square variate of 1 d.f. If we can make the appropriate orthogonal transformation to express the second and third terms of (vi) as Chi-Square variates of 1 d.f. included in the first, the expression on the left is therefore a Chi-Square variate of  $(c - 2)$  d.f. What follows is an outline of the proof. We first put, as in 16.02 and 16.04:

$$\sum_{i=1}^{i=c} \frac{p_i \cdot M_{e.ic}^2}{\sigma_e^2} = \sum_{i=1}^{i=c} u_i^2 = \sum_{i=1}^{i=c} \frac{M_{e.ic}^2}{\sigma_{m.i}^2} \quad \dots \quad (\text{vii})$$

$$u_i = \sum_{i=1}^{i=c} \frac{C_i \cdot M_{e.ic}}{\sigma_{m.i}} = \sum_{i=1}^{i=c} \frac{\sqrt{p_i} \cdot C_i \cdot M_{e.ic}}{\sigma_e} \quad \dots \quad (\text{viii})$$

We may now put

$$u_1 = \frac{M_{e.cs}}{\sigma_m} = \frac{\sqrt{p} \cdot M_{e.cs}}{\sigma_e} = \sum_{i=1}^{i=c} \frac{p_i}{\sqrt{p}} \frac{M_{e.ic}}{\sigma_e} = \sum_{i=1}^{i=c} \left(\frac{p_i}{p}\right)^{\frac{1}{2}} \frac{M_{e.ic}}{\sigma_{m.i}} \quad (\text{ix})$$

Thus  $u_1$  is a linear function of the standard  $e$ -score means in accordance with (viii) and satisfies the orthogonal condition that each  $u$ -score is a score of unit variance since the sum of the square of the linear constants  $(p_i \div p)^{\frac{1}{2}}$  is unity. In virtue of (vi), we may also put

$$u_2 = \frac{(k_{ba.cs} - k_{ba})\sqrt{p \cdot V_{a.s}}}{\sigma_e} = \sum_{i=1}^{i=c} p_i \frac{A_i}{\sqrt{p \cdot V_{a.s}}} \cdot \frac{M_{e.ic}}{\sigma_e},$$

$$\therefore u_2 = \sum_{i=1}^{i=c} \frac{(p_i)^{\frac{1}{2}} A_i}{\sqrt{p \cdot V_{a.s}}} \frac{M_{e.ic}}{\sigma_{m.i}} \quad \dots \quad (\text{x})$$

Thus  $u_2$  is a linear function of the standard  $e$ -score means and satisfies the orthogonal condition

$$\sum_{i=1}^{i=c} \frac{p_i \cdot A_i^2}{p \cdot V_{a.s}} = 1.$$

The definition of  $u_1$  and  $u_2$  is also consistent with the essential orthogonal condition that the sum of the cross products of the linear coefficients vanishes, since this sum is

$$\sum_{i=1}^{i=c} \left(\frac{p_i}{p}\right)^{\frac{1}{2}} \frac{p_i^{\frac{1}{2}} \cdot A_i}{\sqrt{p \cdot V_{a.s}}} = \frac{1}{p\sqrt{V_{a.s}}} \sum_{i=1}^{i=c} p_i \cdot A_i = 0.$$





For what follows we may write this in the form

$$F_k = \frac{(p-2)(k_{ba.cs} - k_{ba})^2 p \cdot V_{a.s}}{(1 - r_{ab.cs}^2) p \cdot V_{b.cs}} \quad . \quad . \quad . \quad (xv)$$

All the essential relations for a significance test of the departure of the regression coefficient from its expected value are implicit in the foregoing derivation of  $F_{cu}$ . We first adapt (iii) above as follows:

$$\begin{aligned} (1 - r_{ab.cs}^2) V_{b.cs} &= E_a E_{b.a} (x_{b.a} - M_{b.as})^2 - (M_{b.cs} - M_{b.s})^2 - (k_{ba.cs} - k_{ba})^2 V_{a.s} \\ &= E_a E_{b.a} (e_{b.a}^2) - M_{e.cs}^2 - (k_{ba.cs} - k_{ba})^2 V_{a.s}, \\ \therefore \frac{(1 - r_{ab.cs}^2) p \cdot V_{b.cs}}{\sigma_e^2} &= \frac{\sum_{i=1}^{i=c} \sum_{j=1}^{j=p_i} e_{j.cs}^2}{\sigma_e^2} - \frac{M_{e.cs}^2}{\sigma_m^2} - \frac{(k_{ba.cs} - k_{ba})^2 p \cdot V_{a.s}}{\sigma_e^2}. \end{aligned}$$

The first term on the right is a Chi-Square variate of  $p$  degrees of freedom, and each of the remaining terms is a Chi-Square variate of 1 d.f. Moreover, we have shown above that we can express the two latter as square  $w$ -scores included in the  $p$ -fold sum of square  $w$ -scores equivalent to the first term. Thus the expression on the left of the equation is a Chi-Square variate of  $(p-2)$  degrees of freedom; and the standardised sum of squares in the numerator of  $F_k$  in (xv) is equivalent to a  $w$ -score we have eliminated from the denominator. In short,  $F_k$  in (xv) is the ratio of a Chi-Square variate of 1 d.f. to a Chi-Square variate of  $(p-2)$  d.f. defined more compactly as

$$F_k = \frac{(p-2)(k_{ba.cs} - k_{ba})^2 V_{a.s}}{(1 - r_{ab.cs}^2) V_{b.cs}} \quad . \quad . \quad . \quad (xvi)$$

This ratio defines the distribution of the deviation of the regression coefficient from its expected value whether the latter does or does not numerically exceed zero. If  $k_{ba} = 0$  it is identical with  $F_{re}$  in (i); but we cannot otherwise express the distribution of the product moment index as a Type VI variate. In fact, of course, an exact test for zero covariance is redundant, since it suffices

(a) to test first whether  $\eta_{ba}^2$  significantly exceeds zero by recourse to the ratio denoted by  $F_{mu}$  above, as in 16.08;

(b) to test subsequently whether regression is linear by recourse to  $F_{cu}$  in (xiv).

The reader will note that the square root of  $F_k$  as defined above is a  $t$ -variate of  $(p-2)$  degrees of freedom, since the numerator has only 1 d.f. Thus we may use the  $t$ -table. Similarly,  $F_{re}$  in (i) is a  $t$ -variate; and we may test for zero covariance, the appropriate  $t$ -ratio being

$$t = r \sqrt{\frac{p-2}{1-r^2}} \quad . \quad . \quad . \quad . \quad (xvii)$$

The derivation of a so-called exact test corresponding to (xxii) of 17.03 introduces no new issue of principle. If we have two independent samples of  $p_1$  and  $p_2$  paired scores respectively, we shall have two estimates of residual variance ( $\sigma_e^2$ ), viz.:

$$R_1 = \frac{p_1(1 - r_{ab.1}^2) V_{b.1}}{p_1 - 2}; \quad R_2 = \frac{p_2(1 - r_{ab.2}^2) V_{b.2}}{p_2 - 2}.$$

Hence we may test whether the residual variation is the same in both samples by the variance ( $F$ ) ratio

$$F_R = \frac{R_1}{R_2} \quad . \quad . \quad . \quad . \quad . \quad . \quad (xviii)$$



If this ratio does not exceed its expected value unduly, i.e. if it satisfies what criterion of significance we agree to adopt, we may proceed as in 13.05 and 16.07 basing our estimate of the residual variance on the *mean*. Thus we may write

$$E_s(1 - r_{ab \cdot 1}^2)V_{b \cdot 1} = \frac{p_1 - 2}{p_1} \sigma_e^2; \quad E_s(1 - r_{ab \cdot 2}^2)V_{b \cdot 2} = \frac{p_2 - 2}{p_2} \sigma_e^2;$$

$$E_s \cdot M(V_b) = \sigma_e^2 \left( \frac{p_1}{p_1 + p_2} \cdot \frac{p_1 - 2}{p_1} + \frac{p_2}{p_1 + p_2} \cdot \frac{p_2 - 2}{p_2} \right)$$

$$= \left( \frac{p_1 + p_2 - 4}{p_1 + p_2} \right) \sigma_e^2.$$

Accordingly we may define a statistic by the relations

$$E_s(s_{e \cdot c}^2) = \sigma_e^2 \quad \text{and} \quad s_{e \cdot c}^2 = \left( \frac{p_1 + p_2}{p_1 + p_2 - 4} \right) M(V_b),$$

$$\therefore s_{e \cdot c}^2 = \frac{p_1(1 - r_{ab \cdot 1}^2)V_{b \cdot 1} + p_2(1 - r_{ab \cdot 2}^2)V_{b \cdot 2}}{p_1 + p_2 - 4} \quad \text{(xix)}$$

The variance ( $\sigma_{k \cdot c}^2$ ) of the distribution of the difference ( $k_{ab \cdot 1} - k_{ab \cdot 2}$ ) is the sum of the variances of the distributions of  $k_{ab \cdot 1}$  and  $k_{ab \cdot 2}$ , i.e.

$$\sigma_{k \cdot c}^2 = \sigma_{k \cdot 1}^2 + \sigma_{k \cdot 2}^2 = \sigma_e^2 \left( \frac{1}{p_1 \cdot V_{a \cdot 1}} + \frac{1}{p_2 \cdot V_{a \cdot 2}} \right).$$

Hence we may take as our unbiased estimate of  $\sigma_{k \cdot c}^2$

$$s_{k \cdot c}^2 = s_{e \cdot c}^2 \left( \frac{1}{p_1 \cdot V_{a \cdot 1}} + \frac{1}{p_2 \cdot V_{a \cdot 2}} \right)$$

We thus obtain a *t*-ratio of  $(p_1 + p_2 - 4)$  degrees of freedom:

$$t_c = \frac{(k_{ab \cdot 1} - k_{ab \cdot 2})}{s_{k \cdot c}} \quad \text{(xx)}$$

We have still to dispose of an issue mentioned in the opening paragraph of this section, *viz.* what is the probability that a particular observation will exceed its estimated value given by the regression line? We can set approximate confidence limits to the regression score, if we assume that the distribution of the *e*-scores is normal. The deviation of an observed value of the *B*-score from the regression estimate is by definition  $(x_{b \cdot a} - x_{r \cdot a cs}) = (x_{b \cdot a} - M_{b \cdot cs} - k_{ba \cdot cs} \cdot X_{a \cdot s})$ . For the *same fixed* value of the *A*-score within the fixed-*A* set the expected value of this is

$$E_c(x_{b \cdot a} - M_{b \cdot cs} - k_{ba \cdot cs} \cdot X_{a \cdot s}) = M_{b \cdot as} - M_{b \cdot s} - k_{ba \cdot s} \cdot X_{a \cdot s} = 0.$$

We may thus write the deviation of  $(x_{b \cdot a} - x_{r \cdot a cs})$  from its expected value as

$$(x_{b \cdot a} - M_{b \cdot a}) - (M_{b \cdot cs} - M_b) - (k_{ba \cdot cs} - k_{ba})X_{a \cdot s} = e_{b \cdot a} - M_{e \cdot cs} - (k_{ba \cdot cs} - k_{ba})X_{a \cdot s}.$$

We have shown above that the squares of each term in this expression expressed in standard form are independent Chi-Square variates, if the distribution of the *e*-scores is normal. Consequently, we may regard it as the sum of independent components on that assumption; and since the variance of the distribution of a raw-score deviation from its mean is necessarily that of the score itself, we may regard the variance of  $(x_{b \cdot a} - x_{r \cdot a cs})$  for a fixed *A*-score as the

sum of 3 additive components. If the estimate  $k_{ba.cs}$  (and hence the value of  $x_{r.acs}$  and  $M_{b.cs}$ ) is referable to  $p$  paired scores, components are

Component	Variance
$\epsilon_{b.a}$	$\sigma_e^2$
$M_{e.cs}$	$\sigma_e^2 \div p$
$(k_{ba.cs} - k_{ba})X_{a.s}$	$X_{a.s}^2 \sigma_{k.s}^2 = (X_{a.s}^2 \cdot \sigma_e^2) \div (p \cdot V_{a.s})$

If we write the variance of  $(x_{b.a} - x_{r.acs})$  as  $\sigma_{r.a}^2$ , we thus have

$$\sigma_{r.a}^2 = \sigma_e^2 \left( 1 + \frac{1}{p} + \frac{X_{a.s}^2}{p \cdot V_{a.s}} \right) \quad \text{. . . . . (xxi)}$$

For computation it is more convenient to write this as

$$\sigma_{r.a}^2 = \sigma_e^2 \left[ 1 + \frac{1}{p} + \frac{A_i^2}{\sum_{i=1}^{i=c} p_i A_i^2} \right]$$

Since this expression involves  $V_{a.s}$ , it presupposes the Model I approach, i.e. sampling with the sub-universe of the fixed- $A$  set. If the  $e$ -score is normal, as we also assume in this context, the deviation  $(x_{b.a} - x_{r.acs})$  involves the differences of independent normal variates, being therefore itself a normal variate. Thus a deviation  $(x_{b.a} - x_{r.acs})$  from its expected value if as great as  $2\sigma_{r.a}$  will occur about 1 in 20 observations in the long run. Actually, we cannot assign an exact value to  $\sigma_e^2$ , and must use our unbiased estimate  $s_e^2$  of (xvi) in 17.03. For an approximate normal test (when  $p$  is large) the appropriate square  $c$ -ratio is therefore

$$\frac{(x_{b.a} - x_{r.acs})^2}{s_e^2 \left[ 1 + \frac{1}{p} + \frac{X_{a.s}^2}{p \cdot V_{a.s}} \right]} = F_e \quad \text{. . . . . (xxii)}$$

By hypothesis the expected value of the numerator in the above is zero, and  $F_e$ , being the ratio of a square standard normal score to the unbiased estimate of its variance, is a  $t$  variate of  $(p - 2)$  degrees of freedom.

### 17.05 THE METHOD OF LEAST SQUARES

When we speak of a sample statistic such as  $k_{ba.cs}$  defined by (i) of 17.02 as the *best estimate* of a parameter (e.g.  $k_{ba}$  in the notation of 17.01) of a universe, we may mean that it satisfies either or both of two criteria: (a) lack of bias; (b) efficiency. An unbiased estimate is a sample statistic whose long run mean value, i.e. mean value for an indefinitely large number of independent samples, is exactly equal to the corresponding universe parameter. We have seen (p. 725) why  $k_{ba.cs}$  defined by (i) of 17.02 is in fact an unbiased estimate of  $k_{ba}$ ; and we shall now ask whether it is the most efficient one.

We have had occasion to refer elsewhere to the concept of efficiency, but have hitherto formulated no general procedure for defining a sample statistic with due regard thereto. In this context two results established in 17.03 simplify our task. We have seen that the distribution of the regression coefficient is normal if: (a) regression is linear; (b) the distribution of errors is normal. When the sample distribution of an estimate is normal, it is possible to define in simple terms a criterion of its statistical efficiency. We speak of one estimate as more efficient than another if we can assert with equal confidence that the true value lies within a smaller range



of values. When the distribution of the estimate, in this context  $k_{ba.cs}$ , is normal with variance  $\sigma_k^2$ , the corresponding standard score of unit variance is  $(k_{ba.cs} - k_{ba}) \div \sigma_k$ ; and we can assert with 95 per cent. confidence that  $k_{ba}$  lies within the range  $k_{ba.cs} \pm 2\sigma_k$  as indicated in 16.05 on p. 696. To make the confidence range of  $k_{ba}$  as small as possible we therefore have to define  $k_{ba.cs}$  in such a way that  $\sigma_k^2$  is smaller than the variance of any alternative normally distributed estimate of  $k_{ba}$ .

The principle of minimal variance last stated is another name for what has long been in use among physicists as a curve fitting device under a different name. The so-called *method of least squares* often invoked to introduce and to justify the use of the sample statistic defined by (i) as an estimate of the constant of a physical law expressed in linear form does in fact justify the assertion (Appendix II) that it is an unbiased one, as we have seen to be true (17.01) for other reasons. We shall now see that the statistic so computed has maximal efficiency, i.e. that  $k_{ba.cs}$  so defined has minimal variance, if we assume a *normal distribution of errors*.

Actually, we do not know the exact value of  $\sigma_k^2$ , but we can regard the ratio of  $(k_{ba.c} - k_{ba})$  to its unbiased estimate  $s_k^2$  as a  $t$ -variate. Whence our problem is to define  $k_{ba.cs}$  so that  $s_k^2$  is a minimum. In (xvii) of 17.03 we have exhibited  $s_k^2$  as a linear function of  $s_e^2$  within the fixed  $A$ -set. Whence it suffices to define  $k_{ba.cs}$  so that  $s_e^2$  is a minimum. We define the term  $x_{r.acs}$  in the expression for  $s_e^2$  to be a point on a line of which the equation is

$$x_{r.acs} = M_{b.cs} + k_{ba.cs} \cdot X_{a.s}.$$

If we use  $E \equiv E_{a.s} \cdot E_{b.a}$ , for brevity

$$E(x_{b.a} - x_{r.acs})^2 = E(x_{b.a} - M_{b.cs})^2 - 2k_{ba.cs} \cdot E(x_{b.a} - M_{b.cs})X_{a.s} + k_{ba.cs}^2 \cdot E(X_{a.s}^2).$$

Whence from (xv) in 17.03:

$$\frac{p-2}{p} s_e^2 = V_{b.cs} - 2k_{ba.cs} \text{Cov}(x_{a.cs}, x_{b.cs}) + k_{ba.cs}^2 \cdot V_{a.s}.$$

In this expression  $\text{Cov}(x_{a.cs}, x_{b.cs})$  is the sample covariance of the  $A$ -scores and the  $B$ -scores. If we are now to define  $k_{ba.cs}$  in such a way that  $s_e^2$  is a minimum, we must put

$$\begin{aligned} \frac{ds_e^2}{dk_{ba.cs}} &= 0, \\ \therefore -2 \text{Cov}(x_{a.cs}, x_{b.cs}) + 2k_{ba.cs} \cdot V_{a.s} &= 0, \\ \therefore k_{ba.cs} &= \frac{\text{Cov}(x_{a.cs}, x_{b.cs})}{V_{a.s}}. \end{aligned}$$

In defining  $k_{ba.cs}$  in such a way that it is an unbiased estimate of  $k_{ba}$ , as shown in 17.01, we have thus defined it so that  $s_e^2$  and  $s_k^2$  is a minimum. Since we can express  $(k_{ba.cs} - k_{ba})$  in terms of  $s_k^2$  as a  $t$ -variate, we have therefore so defined it as to make its confidence range as parsimonious as possible.

## 17.06 REGRESSION IN THE DOMAIN OF CONCURRENCE

In statistical enquiries it may happen that observational data involving two variates, e.g. family income and sickness rates of mothers, appear to cluster near a straight line when plotted on graph paper. It is then possible to assign by the foregoing procedure a straight line of best fit for the regression of one variate on the other, e.g. mother's sickness rate on family income.

It is customary to speak of the equation definitive of the fitted line as a regression equation, and to regard it as a device for predicting the value of one variate when we know the other. Needless to say, prediction in this context means at best assigning confidence limits to a so-called expected value of the variate; but the legitimacy of doing so raises issues quite outside the scope of considerations which justify the argument of 17.02–17.03 in the domain of physical laws.

As we have now seen, the procedure we call the derivation of a regression line is what physicists call the least square method of determining the best value of a physical constant; and its rationale in this context of physical laws implicitly signifies what we have elsewhere designated a *consequential* relationship. With Pearson's collaboration, Galton, to whom the term regression is due, applied it to such situations as the concomitant variation of physical measurements of relatives, e.g. when one plots the height of one member of a twin pair against the height of the other; but in such situations the relationship involved is *concurrent* (*vide* 8.01 in Vol. I), and it is by no means clear that mathematical assumptions appropriate to a theoretical analysis of the sampling process in the consequential domain of a physical law are as relevant to concurrent relationships as Pearson believed.

A paradox which confronts the student in a different context may serve to focus attention on the need to scrutinise such assumptions when we transfer them to the domain of concurrence. When the relationship under discussion is consequential, the square of the correlation coefficient is a precise measure of *explained variance* defined by the relation  $\sigma_t^2 = r_{ab}^2 \cdot \sigma_b^2$  of (xvi) in 17.01; but this is not a rule universally applicable to situations in which linear regression arises. This we shall see more fully in the next chapter. Here we may dispel a difficulty which otherwise confronts the student, if we anticipate a conclusion established later, when we derive (p. 791) the correlation coefficient of two tests *A* and *B* as  $r_{ab} = a_u b_u$  in terms of their so-called communalities  $a_u^2$  and  $b_u^2$ . Hence for two tests with the same communality  $r_{ab} = a_u^2$  which is the fraction of the variance of the test score distribution attributable to a component common to each set. Thus we identify the explained fraction of variance with the correlation coefficient itself in contradistinction to its square.

We can get some light on this seeming inconsistency, if we recall the simplest form of the umpire bonus model of Chapter 9 in Vol. I, the score system being

$$x_a = x_u + x_{a.o}; \quad x_b = x_u + x_{b.o}.$$

In the *consequential* domain of the relation between the player's score and that of the umpire we then have

$$r_{au} = \frac{\sigma_u}{\sigma_a}; \quad r_{bu} = \frac{\sigma_u}{\sigma_b}.$$

In the concurrent domain of the two players' scores we have

$$r_{ab} = \frac{\sigma_u^2}{\sigma_a \sigma_b} = r_{au} \cdot r_{bu}.$$

If both players toss the same die the same number of times  $\sigma_a = \sigma_b$ , so that  $r_{au} = r_{bu}$  and

$$r_{bu}^2 = r_{ab} = r_{au}^2$$

To the present writer, it seems that this distinction resolves the paradox under discussion when we consider the way in which we derive a line of best fit by the method of the last two sections. In plotting the results of a physical experiment we may distinguish between two procedures: (a) each value of the so-called *dependent* variate, e.g. the stretch of a spring, plotted against a particular value of the other variate may truly correspond to one value of the latter, as when we successively measure the stretch produced by adding one and the same load to the



scale pan; (b) each value of the dependent variate (e.g. blood sugar) plotted against one and the same value of the other variate (e.g. insulin dosage) involves an *unacknowledged* error of observation in the measurement of the latter. Either way, the customary procedure in the conduct of an experiment entails what we have tacitly assumed in fitting a line to our observations by the method of least squares, *viz.* that *all the errors of observation arise in assigning a value to the so-called dependent variate.*

In terms of our model situation, we may therefore say that we treat the situation as a player-umpire relation whether our laboratory procedure does (b) or does not (a) involve concurrent liability of both variates to error. If, in reality, both variates of an experimental set-up are subject to error of observation, our method of plotting our observations transfers errors of one sort to the opposite side of the balance sheet, as if we were to assign to player B (the dependent variate) the score  $x_b = x_u + x_{a.o} + x_{b.o}$  and the score  $x_u = x_a$  to player A in the umpire-bonus set-up.

In theory  $k_{ab}$  is the reciprocal of  $k_{ba}$  when there is no error variance, i.e. when  $r_{ab} = 1$ , since in that event

$$k_{ab} = r_{ab} \frac{\sigma_a}{\sigma_b} = \frac{\sigma_a}{\sigma_b}; \quad k_{ba} = r_{ab} \frac{\sigma_b}{\sigma_a} = \frac{\sigma_b}{\sigma_a} \quad . \quad . \quad . \quad . \quad (1)$$

In laboratory practice, of course, this is not so; but the fact that application of the method of 17.02 leads to two different lines of best fit does not constitute a dilemma. In the laboratory there is commonly a clear-cut operational distinction between the variate we deem to be dependent (e.g. volume) and the alternative one, i.e. the one which is more amenable to direct control (e.g. pressure). In applying the method of 17.02 to laboratory data we do not then have to make a choice between two ways of fitting a line. Admittedly, this is not always so. The laboratory worker may be free to choose one of two procedures: (a) to measure the stimulus requisite to produce a fixed response; (b) to measure the response evoked by a fixed stimulus. In either case, however, fixing the value of the so-called independent variable may in fact be subject to experimental error, neglected by the way we plot our data. In terms of the allocation of errors to one or other side of the balance sheet, the two procedures are not identical.

In laboratory enquiry, the very fact that one variable is under the control of the investigator signifies that the relation sought is consequential. On the other hand, statistical enquiries in the domain of sociology, psychology and biology commonly confront us with *concurrent* relationships of which the common element is not under control. The end in view may decide the proper choice of one or other variate as dependent, i.e. the variate  $x_b$  when we speak of the regression of  $x_b$  on  $x_a$ ; but what legitimate aims we may indeed pursue raises issues foreign to the considerations which commend the methods of 17.03–17.04 in the domain of experiment. If we plot weights of schoolboys against age (or *vice versa*), we may adopt one of two procedures. In these days of computing machines, it is common practice to determine a value of  $Cov(x_a, x_b)$  based on the cross-products of *all* the scores, and it is no longer clear that we have to conceive the sampling process as restricted to the sub-universe of the fixed-*A* set or of the fixed-*B* set. Alternatively, we may group all children of over 8 years and no more than  $8\frac{1}{2}$ , labelling the age of the group as  $8\frac{1}{4}$  years. Our calculated  $r_{ab}$  will then be based on cross-products of the *B*-scores (weights) and the corresponding *fixed* age group medians. This procedure is superficially more like laboratory procedure than is the alternative; but the likeness holds good only in the domain of arithmetic.

The implications of the use of curve-fitting by least squares do not admit of any formidable ambiguities in laboratory practice ; and if we fully understand the implicit, as well as the explicit, assumptions we make when we use the methods of 17.03–17.04 for the analysis of experimental



data, we shall avoid the pitfalls which beset us when we use the technique of regression in statistical enquiries outside the laboratory. Contrariwise, a too facile view of the similarities between the two situations will assuredly lead us astray. At the start, we should be clear about the implications of the fact that the method of least squares transplanted into the field of biology and sociology by Pearson was originally a *theory of error*, a basic assumption being that the physicist can control every relevant variable in an experimental set-up other than variation arising from unreliability of his recording apparatus, such variation being free from systematic bias. In biological experiment, one has commonly to take stock of individual variation, e.g. with respect to genetic constitution; but the investigator, with a justifiable intention of propounding a law, implicitly assumes the possibility of repeating observations based on different individuals without introducing a systematic source of variability.

In fact, we assume more than this when we invoke statistical tests dealt with in this context. Our postulate is that the source of residual variation is the same for all samples; and this makes the *Principle of the Fixed-A set* the king-pin of our theoretical edifice. The postulate itself is admissible in comparison of physical experiments in which investigators of equal competence employ the same instruments or instruments of equal precision; but we shall shun the temptation to regard statistics as an efficacious remedy for shoddy experiments in the biological domain, if we are alert to the need for factual support to sustain the proposition that the non-systematic components of variation in different samples of living creatures are necessarily equivalent. Only the strictest attention to selection of stocks standardised with respect both to nature and to nurture, age and season, can confer plausibility of any such assumption implicit in what Churchill Eisenhart calls the Model I approach. The admissible postulates of physical experiment, and those the biologist may be able to adopt with justifiable confidence on that understanding, are at least open to grave doubt in many situations which prompt sociologists and psychologists to employ regression equations. In such enquiries, what is usually a more important source of variation is a complex of external agencies we have no power to control. Were it otherwise, our residual variance would be simply a measure of the failure of our powers of observation to detect a law of nature. As it is, our residual variance is to no small extent a record of the inadequacy of any simple law as a valid description of our observations, and an admission of our powerlessness to recreate a *unique historic event*.

To make the last assertion more tangible let us recall the law of the stretched spring. When we state such a law, the end in view is to tell us by how much we can extend a spring, if we measure the extension with sufficient accuracy under specified loads. A latent assumption is that our laboratory is static. The results would indeed be different if we made our observations in an aeroplane at different (and unknown) heights above sea level in virtue of variations w.r.t. the gravitational constant  $g$ . The best we could then hope for is that we could distribute our observations on the stretch with respect to a specified tension so that differences with respect to elevation would be uniformly distributed. Even in the absence of error inherent in the technique of observation as such, our line of best fit could then tally with the one definitive of the physical law of the static laboratory only in so far as it described the trend of averages. Figuratively speaking, the laboratory of the social scientist and of the vital statistician is always an aeroplane of unknown and changing height above sea level. Errors of observation in the Gaussian sense may be, and indeed commonly are, trivial components of the residual variation undetermined by the course of the regression line.

If it is important on this account to recognise that we cannot rightly equate the *residual* variation of the sociologist or of the vital statistician to instrumental or personal errors of observation as in experimental science, it is no less important to recognise that any statement of a scientific law is complete only in so far as it implies a specification of its own limitations. The laboratory



worker familiar with such limitations can commonly shirk the obligation to make them explicit without compromising the usefulness of conclusions drawn from the law itself. Thus we can safely use an equation prescribing how the density of water varies in relation to temperature at 760 mm. atmospheric pressure without incurring the temptation to invoke its aid to prescribe the density of steam at 120° C. and sea-level pressure. We learn at school that Hooke's law breaks down if the extension approaches breaking point, and that Van der Waals' equation has to replace Boyle's simpler and for most purposes good enough rule in the neighbourhood of absolute zero or of the critical pressure. The *explicit* algebraic formulation of a physical law is always incomplete from this viewpoint, but the experimentalist translates it in action with the reservation that the correct interpretation carries with it a supplementary specification of the boundary conditions of its validity. To say this is to say that the legitimate use of an equation definitive of a structural law in physics lies within the domain of *interpolation*; and the teaching of elementary physics familiarises us with the absurdities which arise when we use it for *extrapolation* beyond the boundaries of its applicability.

This is indeed precisely comparable to what we do, if we succumb to the temptation of using a regression equation as a basis for predicting how a wage increase will affect fertility or infantile mortality. What is a sufficiently well recognised truism in experimental science is a *caveat* we too easily ignore in sociology and vital statistics. For instance, we cannot legitimately infer from the regression of completed family size on family income what the completed family size would be, if we stabilised all incomes at a fixed level, thereby changing the framework of conditions in which the regression relation is valid. The statistical literature of the last fifty years abounds with conclusions of this type, though it is easy to detect the fallacy, if we take stock of a fundamental difference between experimental investigation and statistical description.

We have already had occasion to recognise that there is a clear-cut distinction in experimental science between what we commonly call the dependent and independent, or as we might more informatively say *consequent* and *antecedent* variates. The antecedent (so-called independent) is the one which the investigator has under his direct and deliberate control; and commonly, though not always, it is the only one within his power to control with ease. For instance, we cannot fill a hypodermic syringe with adrenalin by raising the blood pressure of the patient; but one can raise the blood pressure of the patient by injection of the contents of a syringe containing adrenalin.

Now we recognise a relationship as consequential because, and only because, we are able to interfere actively with the course of events; but we are not recording the result of any such active interference when we plot a regression graph of completed family size or maternal morbidity on family income. At least as likely as not, the relationship involved is concurrent; and our plotted data cannot give us any assurance to the contrary. The algebraic treatment of correlation in Chapter 12, in contradistinction to the more customary geometrical approach, can indeed make this logical distinction explicit. We can influence the score of player A, if we record wrongly the result of the umpire's score; but we cannot do so by recording the score of player B wrongly.

In this context, however, a factual as opposed to a schematic illustration may prove more helpful. We may imagine a situation not uncommon in Asia or Africa, *viz.* a population subject to malaria spread over a dry hillside and swampy lowlands around it, the more prosperous Herrenvolk householders having settled on the heights. In the nature of the case, we should then expect to find a correlation between mean income and malaria incidence in the various precincts, and it might well happen that we could plot our statistics as a linear regression graph. In this set-up raising the income of the less prosperous sections of the community might permit more migration from the swampy lowlands and hence less risk of malaria, but only if there were still land available for building on the uplands and only if there were no commensurate



increase in the value of house property. In the absence of any information about the availability of alternative accommodation and the prospects of the building market we therefore lack sufficient reason for inferring what effect an all-round increase of income would have. Since our regression equation contains no information of this sort, it cannot legitimately lead us to forecast the effects of income change.

We may now sum up as follows what we can legitimately mean by prediction in descriptive statistical enquiries :

- (a) a regression graph specifies a sub-sample representative value which we can circumscribe by confidence limits by the methods of 17.03–17.04 on the assumption of a normal error distribution ;
- (b) of itself, the regression equation implies no information concerning the causal relation between the variates, and does not entitle us to make assertions concerning the results of human interference ;
- (c) even if we have additional sources of information to identify the relationship of the variates as one of antecedent to consequent, it is still necessary to remember that :
  - (i) a regression equation describes occurrences in a specified framework of repetition ;
  - (ii) assertions concerning the effects of human interference will not necessarily be true if the latter prescribes a different framework.

In what we have discussed so far, all the emphasis has been on the distinction between the consequential domain of experiment and a concurrent domain which is amenable only to passive observation. The distinction has an implication which is worthy of more explicit comment from a viewpoint adumbrated in a passing remark to the effect that every sociological situation is a *unique historical event*. In the derivation of the significance tests of this chapter, we have assumed what we here call the principle of the fixed- $A$  set. In other words, we view the situation from the viewpoint of what Churchill Eisenhart calls Model I, i.e. as one we can repeat at will in the same way. In a well-controlled laboratory set-up this is a meaningful assumption. It is at least permissible to doubt whether it has any meaning whatsoever in the domain of sociology. Admittedly, it will have one for those who can stomach Plato's conception that the shadow world of human experience is but a sample from the infinite and eternally repetitious universe of universals. To others, its semantic credentials will be less patent.

#### 17.07 PARTIAL REGRESSION AND MULTIPLE CORRELATION

Hitherto we have confined our attention to regression as a linear relation between two variates. Perfect linear regression of the  $B$ -score ( $x_b$ ) on the  $A$ -score in a bivariate universe signifies that the mean  $B$ -score ( $M_{b \cdot a}$ ) associated with a particular  $A$ -score ( $x_a$ ) is directly proportional to the latter, i.e.

$$M_{b \cdot a} = k_{ba} \cdot x_a + C \quad \text{or} \quad M_{b \cdot a} - M_b = k_{ba} \cdot X_a.$$

If this is so, certain identities follow as tautologies of the grid which summarises the structure of the universe, in particular

$$\text{Cov}(x_a, x_b) = k_{ba} \cdot V_a \quad \text{and} \quad V(M_{b \cdot a}) = k_{ba}^2 \cdot V_a.$$

In random sampling from the bivariate universe of the consequential domain, the unbiased and most efficient estimate ( $k_{ba \cdot cs}$ ) of  $k_{ba}$  has the same relation to the sample covariance and sample  $A$ -score variance, viz. :

$$k_{ba \cdot cs} = \frac{\text{Cov}(x_{a \cdot cs}, x_{b \cdot cs})}{V_{a \cdot cs}}.$$



These expressions constitute a particular case of a linear relation involving several variables. For our purpose it will suffice to illustrate the pattern by consideration of the case which arises when we prescribe the mean value ( $M_{c.ab}$ ) of one score ( $x_c$ ) in terms of particular values of two others ( $x_a$  and  $x_b$ ) connected therewith by the linear relation

[illegible]

The mean value of  $M_{c..ab}$  is the mean  $C$ -score value for the particular set of  $A$ -scores and  $B$ -scores which with them constitute the trivariate universe, whence we can eliminate the constant  $K$  in the usual way :

[illegible]

*Tautologies of Multiple Regression.* Equations (i) and (ii) define the relation between the mean  $C$ -score and particular values of the  $A$ -scores and  $B$ -scores of a trivariate universe in which regression of the  $C$ -score on the other two scores is exactly linear. We can visualise such a universe in the idiom of *Chance and Choice* as the long-run result of a game of which the 3 recorded scores are : (i) the player  $C$  ; (ii) an umpire  $A$  ; (iii) an umpire  $B$ , as in the model of 12.01. To make the model as general as possible for our purpose we need not assume that the scores ( $x_a$  and  $x_b$ ) of the umpires are independent. Thus we are free to regard them as correlated in virtue of the contribution of a third umpire to each of them. The rule is that the player adds to his individual (and hence independent) score ( $x_c$ ) some multiple ( $k_{ca \cdot b} = k_a$ ) and ( $k_{cb \cdot a} = k_b$ ) of each of the umpires, so that

$$x_c = x_e + k_a \cdot x_a + k_b \cdot x_b \quad . \quad . \quad . \quad . \quad . \quad . \quad (\text{iii})$$

In terms of deviations from the score component mean values ( $M_c$ ,  $M_e$ ,  $M_a$  and  $M_b$ ) this is equivalent to

[illegible]

Since the individual score ( $x_e$ ) of player  $C$  is independent of the umpire's contribution, it can take any value for a fixed value of the  $A$ -score or the  $B$ -score, so that  $M_{e \cdot ab} = M_e$  being therefore constant and equivalent to  $C$  in (i). Thus regression of the  $C$ -score on the  $A$ -scores of the two umpires is linear.

If we multiply (iv) by  $X_a$  or  $X_b$  and take the mean value of the product, we at once derive as a grid tautology

$$Cov(x_a, x_c) = Cov(x_e, x_a) + k_a \cdot V_a + k_b Cov(x_a, x_b);$$

$$Cov(x_b, x_c) = Cov(x_e, x_b) + k_b \cdot V_b + k_a Cov(x_a, x_b).$$

Since the player's individual score is independent of that of either umpire, the long-run value of  $Cov(x_a, x_e)$  and  $Cov(x_b, x_e)$  is zero and

$$Cov(x_a, x_c) = k_a \cdot V_a + k_b Cov(x_a, x_b) \quad . \quad . \quad . \quad . \quad (v)$$

$$Cov(x_b, x_c) = k_b \cdot V_b + k_a Cov(x_a, x_b) \quad . \quad . \quad . \quad . \quad (vi)$$

We can now eliminate  $k_a$  or  $k_b$ , e.g. if we put

$$Cov(x_a, x_c) Cov(x_a, x_b) = k_a \cdot V_a \cdot Cov(x_a, x_b) + k_b Cov^2(x_a, x_b);$$

$$Cov(x_b, x_c)V_a = k_a \cdot V_a \cdot Cov(x_a, x_b) + k_b \cdot V_a \cdot V_b,$$

$$\therefore k_b = \frac{Cov(x_a, x_c) Cov(x_a, x_b) - V_a Cov(x_b, x_c)}{Cov^2(x_a, x_b) - V_a \cdot V_b} \quad \text{. . . (vii)}$$

Similarly,

$$k_a = \frac{\text{Cov}(x_b, x_c) \text{Cov}(x_a, x_b) - V_b \text{Cov}(x_a, x_c)}{\text{Cov}^2(x_a, x_b) - V_a \cdot V_b} \quad . \quad . \quad . \quad (\text{viii})$$

For convenience at a later stage, we may write these results in terms of sums of square deviations or products. If the 3-dimensional grid contains  $n$  score triplets, we may write

$$S_j = n \cdot V_j = \sum_{j=1}^n X_j^2 \quad \text{and} \quad S_{ij} = \sum_{p=1}^n X_{i.p} \cdot X_{j.p} \quad . \quad . \quad . \quad (\text{ix})$$

The above expression then takes the form

$$k_a = \frac{S_{bc} \cdot S_{ab} - S_{bb} \cdot S_{ac}}{S_{ab}^2 - S_{aa} \cdot S_{bb}} \quad \text{and} \quad k_b = \frac{S_{ac} \cdot S_{ab} - S_{aa} \cdot S_{bc}}{S_{ab}^2 - S_{aa} \cdot S_{bb}} \quad . \quad . \quad (\text{x})$$

The partial regression coefficients  $k_a$  and  $k_b$  are expressible in terms of partial correlation coefficients. For brevity, we may first write

$$K_{ca.b} = k_a \frac{\sigma_a}{\sigma_c} \quad \text{and} \quad K_{cb.a} = k_b \frac{\sigma_b}{\sigma_c} \quad . \quad . \quad . \quad (\text{xi})$$

If we substitute for  $\text{Cov}(x_a, x_b)$  in (vii)–(viii)  $r_{ab} \cdot \sigma_a \cdot \sigma_b$  and *mutatis mutandis* for  $\text{Cov}(x_a, x_c)$ ,  $\text{Cov}(x_b, x_c)$ , we derive

$$K_{ca.b} = \frac{r_{ac} - r_{bc} \cdot r_{ab}}{1 - r_{ab}^2} \quad \text{and} \quad K_{cb.a} = \frac{r_{bc} - r_{ac} \cdot r_{ab}}{1 - r_{ab}^2} \quad . \quad . \quad (\text{xii})$$

So far we have considered the regression of the  $C$ -score on the  $A$ -score and the  $B$ -score, in which case there is no ambiguity in the substitution  $k_a = k_{ca.b}$ . If our concern is with the regression of the  $A$ -score, we need to distinguish  $k_{ca.b}$  from  $k_{ac.b}$  in the corresponding regression equation from which we derive the former in the same way, and write more fully

$$K_{ac.b} = k_{ac.b} \frac{\sigma_a}{\sigma_c} \quad \text{and} \quad K_{ca.b} = k_{ca.b} \frac{\sigma_a}{\sigma_c}.$$

Whence we obtain

$$K_{ac.b} \cdot K_{ca.b} = \frac{(r_{ac} - r_{bc} \cdot r_{ab})^2}{(1 - r_{ab}^2)(1 - r_{bc}^2)} \quad . \quad . \quad . \quad (\text{xiii})$$

Whence from (iv) of 12.08, we get

$$K_{ac.b} \cdot K_{ca.b} = r_{ac.b}^2 = k_{ac.b} \cdot k_{ca.b} \quad . \quad . \quad . \quad (\text{xiv})$$

Similarly, we may define the remaining 4 regression coefficients in terms of  $r_{ab.c}$  and  $r_{bc.a}$ .

The use and build-up of what it is customary to call the *multiple correlation coefficient* is easy to understand if we visualise the 2-dimensional grid of simple linear regression as a *scatter diagram*, i.e. a cloud of points on a graph. A product-moment index approaching unity then signifies that the points cluster closely round the line of best fit defined algebraically by the regression equation. We may express this conception formally in another way, as in the derivation of (xxi) in 11.04. To say that such a line gives a perfect fit to the data means that all the points lie on it; and this signifies a one-to-one correspondence of  $x_b$  to  $M_{b.a}$  for every value of  $x_a$ . If regression of  $x_b$  on  $x_a$  is indeed linear, interchanging the then equally spaced values of  $M_{b.a}$  at the foot of each column of the score-frequency grid with the equally spaced values of the  $x_a$  border-scores at the head of each column, is equivalent to a change of the origin and scale of the  $A$ -score distribution; and we have seen (p. 353, Vol. I) that this does not affect the value of  $r_{ab}$ . In other words, the  $p$ - $m$  correlation ( $r_{bm}$ ) of  $x_b$  with  $M_{b.a}$  is the correlation



$(r_{ab})$  of  $x_b$  with  $x_a$  when regression is truly linear, and its numerical value is a yardstick of the good fit. Formally, we may express the identity thus

$$r_{bm} = \frac{E[X_b(M_{b \cdot a} - M_b)]}{\sqrt{V_b \cdot V(M_{b \cdot a})}}.$$

In this expression linear regression implies (xi) in 11.04, i.e. that  $V(M_{b..a}) = k_{ba}^2 \cdot V_a$  and

$$E[X_b(M_{b \cdot a} - M_b)] = k_{b \cdot a} \cdot E(X_a \cdot X_b) = k_{b \cdot a} \cdot \sqrt{V_a \cdot V_b} \cdot r_{ab},$$

$$\therefore r_{bm} = r_{ab}.$$

These considerations suggest that we may profitably explore the correlation between the actual value of  $X_c$  and  $(M_{c \cdot ab} - M_c)$  in (ii) above as a criterion of satisfactory fit, i.e. how closely particular values of  $X_c$  correspond to corresponding mean values on the assumption that 2 other variates are relevant. Accordingly, we define a multiple correlation coefficient for such a set-up as

$$R_c = \frac{Cov(x_c, M_{c.ab})}{\sqrt{V_c \cdot V(M_{c.ab})}} \quad (xv)$$

In (xv) the value of  $V(M_{c..ab})$  follows from (ii), since

$$\begin{aligned} (M_{c \cdot ab} - M_c)^2 &= k_a^2 \cdot X_a^2 + k_b^2 \cdot X_b^2 + 2k_a \cdot k_b \cdot X_a \cdot X_b, \\ \therefore V(M_{c \cdot ab}) &= k_a^2 \cdot V_a + k_b^2 \cdot V_b + 2k_a \cdot k_b Cov(x_a, x_b) \quad . \quad . \quad (xvi) \end{aligned}$$

Similarly,

$$\begin{aligned} X_c \cdot M_{c \cdot ab} &= k_a(X_c \cdot X_a) + k_b(X_c \cdot X_b) + M_c \cdot X_c, \\ \therefore Cov(x_a, M_{c \cdot ab}) &= k_a Cov(x_a, x_c) + k_b Cov(x_b, x_c) \quad . \quad . \quad . \quad (xvii) \end{aligned}$$

Whence from (v) and (vi)

$$Cov(x_c, M_{c,ab}) = k_a^2 \cdot V_a + k_b^2 \cdot V_b + 2k_a k_b Cov(x_a, x_b).$$

Hence from (xvi)

$$Cov(x_c, M_{c \cdot ab}) = V(M_{c \cdot ab}).$$

Whence from (xv) :

$$\begin{aligned} R_c^2 &= \frac{\text{Cov}(x_e, M_{c.ab})}{V_c} \\ &= \frac{k_a \text{Cov}(x_a, x_e)}{V_c} + \frac{k_b \text{Cov}(x_b, x_e)}{V_c} \\ &= k_a.r_{ac}\frac{\sigma_a}{\sigma_c} + k_b.r_{bc}\frac{\sigma_b}{\sigma_c}, \\ \therefore R_c &= \sqrt{k_a.r_{ac}\frac{\sigma_a}{\sigma_c} + k_b.r_{bc}\frac{\sigma_b}{\sigma_c}} . . . . . (\text{xviii}) \end{aligned}$$

The analogy between the *multiple correlation coefficient* ( $R_c$ ) and the product moment index  $r_{ab}$  of the bivariate universe extends beyond the explanation given above. We may identify the true value of the C-score with the universe mean ( $M_{c \cdot ab}$ ), in which case, (iv) takes the form

$$\begin{aligned} X_c &= M_{c \cdot ab} + X_e, \\ \therefore V_c &= V(M_{c \cdot ab}) + V_e = Cov(x_c, M_{c \cdot ab}) + V_e, \\ \therefore V_c &= R^2 \cdot V_e + V_e, \\ \therefore V_e &= (1 - R^2)V_c. \end{aligned}$$

*The Unbiased Estimates of the Coefficients.* So far we have defined  $k_a$  and  $k_b$  as constants connecting particular values of  $X_a$  and  $X_b$  with the true value of  $(M_{e.ab} - M_b)$ . As with simple linear regression, we can define unbiased estimates of  $k_a$  and  $k_b$  if we look at the problem as a case of Churchill Eisenhart's Model I (p. 548). We conceive that we are repeating observations of the  $C$ -scores on exactly the same set of  $A$ -scores and  $B$ -scores in each experiment of which we have a sample before us. In our model set-up the player's individual score component ( $X_e$ ) now takes the place of the error or residual source of random variation. We shall denote the operation of taking a single  $n$ -fold sample of score triplets by  $E_n(\dots)$  and the expected value of a sample parameter by  $E_s(\dots)$ . Thus

$$\text{Cov}(x_{a.s}, x_{c.s}) = E_n(X_e.X_a) + k_a.V_{a.s} + k_b.\text{Cov}(x_{a.s}, x_{b.s}),$$

$$\therefore E_s.\text{Cov}(x_{a.s}, x_{c.s}) = E_s.E_n(X_e.X_a) + k_a.E_s(V_{a.s}) + k_b.E_s.\text{Cov}(x_{a.s}, x_{b.s}).$$

Now the expected value of  $E_n(X_e.X_a)$  is zero in virtue of the independence of the error component of the  $C$ -score; and  $V_{a.s}$  is constant within the framework of the Model I set-up, as is also  $\text{Cov}(x_a, x_b)$ , so that

$$E_s.\text{Cov}(x_{a.s}, x_{c.s}) = k_a.V_{a.s} + k_b.\text{Cov}(x_{a.s}, x_{b.s}).$$

Similarly,

$$E_s.\text{Cov}(x_{b.s}, x_{c.s}) = k_b.V_{b.s} + k_a.\text{Cov}(x_{a.s}, x_{b.s}).$$

By elimination in the usual way we have

$$\begin{aligned} k_a &= \frac{\text{Cov}(x_{a.s}, x_{b.s}) E_s.\text{Cov}(x_{b.s}, x_{c.s}) - V_{b.s} . E_s.\text{Cov}(x_{a.s}, x_{c.s})}{\text{Cov}^2(x_{a.s}, x_{b.s}) - V_{a.s} . V_{b.s}} \\ &= \frac{E_s.\text{Cov}(x_{a.s}, x_{b.s}) \text{Cov}(x_{b.s}, x_{c.s}) - V_{b.s} \text{Cov}(x_{a.s}, x_{c.s})}{\text{Cov}^2(x_{a.s}, x_{b.s}) - V_{a.s} . V_{b.s}}. \end{aligned}$$

We may thus define the statistic which is an unbiased estimate of  $k_a$  within the fixed set of  $A$ -scores and  $B$ -scores by the relations

$$E_s(k_{a.s}) = k_a$$

and

$$k_{a.s} = \frac{\text{Cov}(x_{a.s}, x_{b.s}) . \text{Cov}(x_{b.s}, x_{c.s}) - V_{b.s} \text{Cov}(x_{a.s}, x_{c.s})}{\text{Cov}^2(x_{a.s}, x_{b.s}) - V_{a.s} . V_{b.s}}. \quad (\text{xix})$$

In conformity with our derivation of the most efficient estimate of  $k_{ba}$  in samples from a bivariate universe, we shall suppose that the  $C$ -score is divisible into two components, one ( $x_r$ ) directly proportional to both  $x_a$  and  $x_b$  and the other a residual ( $x_e$ ) whose mean square deviation ( $V_{e.s}$ ) from its sample mean value ( $M_{e.s}$ ) is to be a minimum by appropriate choice of the linear constants  $k_{a.s}$  and  $k_{b.s}$ , defining the relation of  $x_r$  to the  $A$ -score and  $B$ -score. By definition therefore

$$x_c = x_r + x_e = k_{a.s} . x_a + k_{b.s} . x_b + K + x_e \quad . \quad . \quad (\text{xx})$$

We can eliminate the constant  $K$  in the usual way, since

$$M_e = C - k_{a.s} . M_a - k_{b.s} . M_b.$$



Whence we can express (xx) in terms of the deviations of the score components from their mean values as

$$X_c = k_{a.s}.X_a + k_{b.s}.X_b + X_e \quad . \quad . \quad . \quad (xxi)$$

$$\therefore X_e^2 = X_c^2 + k_{a.s}^2 X_a^2 + k_{b.s}^2 X_b^2 - 2k_{a.s} X_a X_c - 2k_{b.s} X_b X_c + 2k_{a.s} k_{b.s} X_a X_b,$$

$$\therefore V_{e.s} = V_{c.s} + k_{a.s}^2 V_{a.s} + k_{b.s}^2 V_{b.s} - 2k_{a.s} \text{Cov}(x_{a.s}, x_{c.s}) \\ - 2k_{b.s} \text{Cov}(x_{b.s}, x_{c.s}) + 2k_{a.s} k_{b.s} \text{Cov}(x_{a.s}, x_{b.s}),$$

$$\therefore \frac{\partial V_{e.s}}{\partial k_{a.s}} = 2k_{a.s} V_{a.s} - 2 \text{Cov}(x_{a.s}, x_{c.s}) + 2k_{b.s} \text{Cov}(x_{a.s}, x_{b.s})$$

and

$$\frac{\partial V_{e.s}}{\partial k_{b.s}} = 2k_{b.s} V_{b.s} - 2 \text{Cov}(x_{b.s}, x_{c.s}) + 2k_{a.s} \text{Cov}(x_{a.s}, x_{b.s}).$$

The condition which makes  $V_{e.s}$  a minimum is that

$$\frac{\partial V_{e.s}}{\partial k_{a.s}} = 0 = \frac{\partial V_{e.s}}{\partial k_{b.s}},$$

$$\therefore k_{a.s}.V_{a.s} = \text{Cov}(x_{a.s}, x_{c.s}) - k_{b.s} \text{Cov}(x_{a.s}, x_{b.s}) \quad . \quad (xxii)$$

and

$$k_{b.s}.V_{b.s} = \text{Cov}(x_{b.s}, x_{c.s}) - k_{a.s} \text{Cov}(x_{a.s}, x_{b.s}) \quad . \quad (xxiii)$$

Our definition of the sample parameters which define the slope of the line of best fit in the sense that the residual variance is minimal thus correspond to the unbiased estimate of the universe parameters  $k_a$  and  $k_b$ .

The student should be able to extend the foregoing derivations to regression involving more than 3 variables. When we have more than 2 regression coefficients to evaluate, it is preferable to solve the basic equations of the form exhibited in (v) and (vi) or (xxii) and (xxiii) by recourse to determinants. In the notation of sums of squares and products, the basic equations for a set-up involving four variables (regression of  $x_d$  on  $x_a$ ,  $x_b$  and  $x_c$ ) take the form \*

$$S_{ad} = k_a.S_{aa} + k_b.S_{ab} + k_c.S_{ac},$$

$$S_{bd} = k_b.S_{bb} + k_a.S_{ab} + k_c.S_{bc},$$

$$S_{cd} = k_c.S_{cc} + k_a.S_{ac} + k_b.S_{bc}.$$

*Numerical Example.* The following are 3 associated variables.

	$x_a$	$x_b$	$x_c$
	5	2	21
	3	4	21
	2	2	15
	4	2	17
	3	3	20
	1	2	13
	8	4	32
Totals	26	19	139
Means	3.714	2.714	19.86

\* For computation short-cuts see Mordecai Ezekiel: *Methods of Correlation Analysis*. Wiley, 1941.

For purposes of computation we may more conveniently work with raw-scores than with score deviations. We then write  $s_i$  as the sum of the score  $x_i$ , so that  $M_i = n \cdot s_i$  and  $s_{ii}$ ,  $s_{ij}$  as the sum of the squares of  $x_i$  and the products  $x_i \cdot x_j$  respectively. For an  $n$ -fold sample we then have

$$S_{ii} = s_{ii} - \frac{s_i^2}{n} \quad \text{and} \quad S_{ij} = s_{ij} - \frac{s_i \cdot s_j}{n} \quad . \quad . \quad . \quad (\text{xxiv})$$

We shall therefore need for the case of 3 variates ( $x_a, x_b, x_c$ ) the following column total raw-scores ( $s_a, s_b, s_c$ ), square ditto ( $s_{aa}, s_{bb}, s_{cc}$ ) and products ( $s_a \cdot s_b, s_a \cdot s_c, s_b \cdot s_c$ ).

Our first step is to tabulate  $s_a, s_b$ , etc., as below :

	$x_a$	$x_b$	$x_c$	$x_a^2$	$x_b^2$	$x_c^2$	$x_a x_b$	$x_a x_c$	$x_b x_c$
	5	2	21	25	4	441	10	105	42
	3	4	21	9	16	441	12	63	84
	2	2	15	4	4	225	4	30	30
	4	2	17	16	4	289	8	68	34
	3	3	20	9	9	400	9	60	60
	1	2	13	1	4	169	2	13	26
	8	4	32	64	16	1024	32	256	128
Total	26	19	139	128	57	2989	77	595	404
	$s_a$	$s_b$	$s_c$	$s_{aa}$	$s_{bb}$	$s_{cc}$	$s_{ab}$	$s_{ac}$	$s_{bc}$

From the above we obtain by recourse to (xxiv)

$$\begin{aligned} S_{ab} &= 77 - \frac{(26 \cdot 19)}{7} = \frac{45}{7}; & S_{aa} &= 128 - \frac{(26)^2}{7} = \frac{220}{7}; \\ S_{ac} &= 595 - \frac{(139 \cdot 26)}{7} = \frac{551}{7}; & S_{bb} &= 57 - \frac{(19)^2}{7} = \frac{38}{7}; \\ S_{bc} &= 404 - \frac{(139 \cdot 19)}{7} = \frac{187}{7}; & S_{cc} &= 2989 - \frac{(139)^2}{7} = \frac{1602}{7}. \end{aligned}$$

We thus derive

$$\begin{aligned} 220k_a + 45k_b &= 551, \\ 45k_a + 38k_b &= 187, \\ k_a &= \frac{12523}{6335} = 1.977, \\ k_b &= \frac{16345}{6335} = 2.580. \end{aligned}$$

Our regression equation for the *predicted* value ( $x_r$ ) of  $x_c$  is thus

$$\left(x_r - \frac{139}{7}\right) = 1.977\left(x_a - \frac{26}{7}\right) + 2.58\left(x_b - \frac{19}{7}\right).$$

If we write as the error  $x_{e.ab}$  of  $x_c$  for a fixed value of  $x_a$  and  $x_b$ , the equation of the  $C$ -score distribution in terms of the estimated value of  $k_a$  and  $k_b$  is

$$\begin{aligned} X_c &= k_a \cdot X_a + k_b \cdot X_b + X_{e.ab}, \\ \therefore X_c^2 &= k_a \cdot X_a \cdot X_c + k_b \cdot X_b \cdot X_c + X_{e.ab} \cdot X_c, \\ \therefore V_c &= k_a \text{Cov}(x_a, x_c) + k_b \text{Cov}(x_b, x_c) + \text{Cov}(x_{e.ab}, x_c). \end{aligned}$$



Alternatively, we may write

$$V_c = k_a^2 \cdot V_a + k_b^2 \cdot V_b + V_e + 2 \text{Cov}(x_a, x_b) + 2 \text{Cov}(x_a, x_e \cdot ab) + 2 \text{Cov}(x_b, x_e \cdot ab).$$

In any case, the fraction of variance explained by the dependence of the  $C$ -score on both the  $A$ -score and the  $B$ -score is

$$\frac{k_a \text{Cov}(x_a, x_c) + k_b \text{Cov}(x_b, x_c)}{V_c} = F_c = \frac{k_a^2 V_a + k_b^2 V_b + 2 \text{Cov}(x_a, x_b)}{V_c}.$$

In this case

$$F_c = 0.981. \dots$$

It is instructive to compare this result with the corresponding measure of explanation calculated on the assumption that the  $A$ -score distribution is the *only* relevant source of systematic variation. Our prediction equation is then

$$x_c - \frac{139}{7} = \left( k_{ca} x_a - \frac{26}{7} \right); k_{ca} = \frac{\text{Cov}(x_a, x_c)}{V_a} = \frac{551}{220} = 2.505.$$

On this assumption

$$F_c = \frac{k_{ca}^2 \cdot V_a}{V_c} = 0.8614.$$

\* \* \* \* \*

*Significance of Individual Variates.* The foregoing numerical example raises the question : how can we decide whether it is advantageous to take stock of an additional variable ? By the method of 17.03 we may obtain 3 independent estimates of error variances for  $n$  triplets on the assumption that variation w.r.t.  $x_a$  and  $x_b$  does *not* contribute to that of  $x_c$ .

$$E_s(K_a \cdot S_{ac}) = \sigma_e^2,$$

$$E_s(K_b \cdot S_{bc}) = \sigma_e^2,$$

$$E_s\left(\frac{S_{cc} - k_a \cdot S_{ac} - k_b \cdot S_{bc}}{n - 3}\right) = \sigma_e^2.$$

Expected values of the statistics  $k_a \cdot S_{ac}$  and  $k_b \cdot S_{bc}$  will exceed  $\sigma_e^2$  if there is significant regression. We thus derive two variance ratios as a basis for the commonly prescribed test of the significance of the contribution of one or other variate :

$$F_a = \frac{(n - 3)K_a \cdot S_{ac}}{S_{cc} - k_a \cdot S_{ac} - k_b \cdot S_{bc}}; F_b = \frac{(n - 3)K_b \cdot S_{bc}}{S_{cc} - k_a \cdot S_{ac} - k_b \cdot S_{bc}}. \quad (\text{xxv})$$

## 17.08 THE DISCRIMINANT FUNCTION

If two classes  $A$  and  $B$  (e.g. males and females) differ w.r.t. several measurable attributes, any one such difference may be absolute in the sense that  $A$ 's measurement is always greater than that of  $B$  (or *vice versa*); but class differences are none the less genuine if expressible only in terms of averages. When this is so a single measurement has little diagnostic value. Thus the fact that the mean height of men is appreciably greater than that of women in the same community does not entitle us to assert with great confidence anything about the sex of an adult whose height is somewhat below the average for females. On the other hand, our assurance would be legitimately greater if we knew that several measurements (e.g. *neck girth*, *hip width*) made on the same individual lay nearer to the female than to the male population mean.

In combining such observations the value of adding a new one will depend partly on whether the sampling variance is small or great and partly on whether it is or is not highly correlated with another already included in the test battery. The fact that some measurements will pay better dividends for diagnostic purposes than others therefore raises the issue: what is the best way of *weighting* each of the observations of the test battery when we combine them in a single index? To get the issue into focus it will suffice to consider a test battery involving only 2 measurements ( $U$  and  $V$ ) as in the accompanying schema:

	Individual Values			Population Means		
	$A$	$B$	Difference	$A$	$B$	Difference
$U$	$x_{j.au}$	$x_{j.bu}$	$d_{j.u} = x_{j.au} - x_{j.bu}$	$M_{u.a}$	$M_{u.b}$	$M_{d.u}$
$V$	$x_{j.av}$	$x_{j.bv}$	$d_{j.v} = x_{j.av} - x_{j.bv}$	$M_{v.a}$	$M_{v.b}$	$M_{d.v}$

If we gave each type of measurement equal weight, the mean values of our diagnostic indices would be  $\frac{1}{2}(M_{u.a} + M_{v.a})$  and  $\frac{1}{2}(M_{u.b} + M_{v.b})$ . Otherwise, we may represent them as

$$I_a = C_u \cdot M_{u.a} + C_v \cdot M_{v.a} \text{ and } I_b = C_u \cdot M_{u.b} + C_v \cdot M_{v.b} \quad . \quad . \quad . \quad (i)$$

For two individuals taken at random, one from each population, the corresponding sample values ( $S_a$  and  $S_b$ ) and their difference ( $D$ ) will then be

$$S_a = C_u \cdot x_{j.au} + C_v \cdot x_{j.av} \text{ and } S_b = C_u \cdot x_{j.bu} + C_v \cdot x_{j.bv} \quad . \quad . \quad . \quad (ii)$$

$$D = S_a - S_b = C_u \cdot d_{j.u} + C_v \cdot d_{j.v}.$$

The mean value of  $D$  is then  $M_d = C_u \cdot M_{d.u} + C_v \cdot M_{d.v}$ . If we assume an approximately normal distribution of individual measurements, and hence of  $D$  itself, we may prefer to define it in such a way as to minimise its variance and hence the limits within which our estimate of  $D$  will lie at a prescribed confidence level. Our problem is then to specify  $C_u$  and  $C_v$  in conformity with this condition. We first note that the partial derivative of the square standard score,

$$\frac{\partial}{\partial C} \cdot \frac{(D - M_d)^2}{V} = \frac{2(D - M_d)}{V} \cdot \frac{\partial D}{\partial C} - \frac{(D - M_d)^2}{V^2} \cdot \frac{\partial V}{\partial C}.$$

The expression on the left vanishes when

$$\frac{\partial D}{\partial C} = \frac{(D - M_d)}{2V} \cdot \frac{\partial V}{\partial C}.$$

Thus to maximise the square standard score of the difference distribution we have to solve two equations:

$$\frac{\partial D}{\partial C_u} = d_{j.u} = \frac{(D - M_d)}{2V} \cdot \frac{\partial V}{\partial C_u};$$

$$\frac{\partial D}{\partial C_v} = d_{j.v} = \frac{(D - M_d)}{2V} \cdot \frac{\partial V}{\partial C_v}.$$

Whence we have

$$d_{j.u} \frac{\partial V}{\partial C_v} = d_{j.v} \cdot \frac{\partial V}{\partial C_u} \quad . \quad . \quad . \quad . \quad . \quad . \quad (iii)$$

Now the variance of the  $D$ -distribution will be the sum of the variances ( $V_{s.a}$  and  $V_{s.b}$ ) of the distributions of  $S_a$  and  $S_b$  in (ii). If  $V_{u.a}$  is the variance of the distribution of  $X_{j.au}$ , etc.:

$$V_{s.a} = C_u^2 \cdot V_{u.a} + C_v^2 \cdot V_{v.a} + 2C_u \cdot C_v \cdot Cov(x_{j.au}, x_{j.av});$$

$$V_{s.b} = C_u^2 \cdot V_{u.b} + C_v^2 \cdot V_{v.b} + 2C_u \cdot C_v \cdot Cov(x_{j.bu}, x_{j.bv}).$$



For brevity we may write

$$V_{u \cdot a} + V_{u \cdot b} = m_{uu}; \quad V_{v \cdot a} + V_{v \cdot b} = m_{vv} \quad . \quad . \quad . \quad (iv)$$

$$Cov(x_{j \cdot au}, x_{j \cdot av}) + Cov(x_{j \cdot bu}, x_{j \cdot bv}) = m_{uv} \quad . \quad . \quad . \quad (v)$$

We then have

$$V = V_{s \cdot a} + V_{s \cdot b} = C_u^2 \cdot m_{uu} + C_v^2 \cdot m_{vv} + 2C_u C_v \cdot m_{uv},$$

$$\therefore \frac{\partial V}{\partial C_u} = 2(C_u \cdot m_{uu} + C_v \cdot m_{uv}) \text{ and } \frac{\partial V}{\partial C_v} = 2(C_v \cdot m_{vv} + C_u \cdot m_{uv}).$$

Whence by substitution in (iii)

$$d_{j \cdot u}(C_v \cdot m_{vv} + C_u \cdot m_{uv}) = d_{j \cdot v}(C_u \cdot m_{uu} + C_v \cdot m_{uv}) \quad . \quad . \quad (vi)$$

Since  $m_{uu}$ ,  $m_{vv}$  and  $m_{uv}$  are parameters of the distributions of measurements, the expressions in parenthesis on each side of (vi) are constants and we may write

$$M_{d \cdot u}(C_v \cdot m_{vv} + C_u \cdot m_{uv}) = M_{d \cdot v}(C_u \cdot m_{uu} + C_v \cdot m_{uv}) \quad . \quad . \quad (vii)$$

If we weight our diagnostic index in the usual way,  $(C_v + C_u) = 1$ . The values of both constants are then obtainable from (vii) in terms of the population mean differences, variances and co-variances. Actually, it is immaterial how we fix one of them, since the multiplication of  $D$  by a fixed constant does not affect the ratio of  $D^2$  to  $V$ . Thus we can write  $C_u = 1$  and solve accordingly.

*Numerical example.* For two measurements each made on 4 males and 4 females, the following will serve :

Individual	$U$		$V$	
	$A$	$B$	$A$	$B$
1	11	12	8	3
2	12	14	12	4
3	14	16	10	2
4	15	18	10	7
Mean	13	15	10	4
Difference	- 2		6	

For this set-up

$$M_{d \cdot u} = -2;$$

$$Cov(x_{u \cdot a}, x_{b \cdot a}) = 1.5;$$

$$V_{u \cdot a} = 2.5;$$

$$V_{v \cdot a} = 2.0;$$

$$m_{uu} = 7.5; \quad m_{uv} = 3$$

$$M_{d \cdot v} = 6;$$

$$Cov(x_{u \cdot b}, x_{v \cdot b}) = 3.5;$$

$$V_{u \cdot b} = 5;$$

$$V_{v \cdot b} = 3.5;$$

$$m_{vv} = 5.5.$$

By substitution in (vii), if we put  $C_u = 1$

$$-2(5.5 C_v + 3) = 6(7.5 + 3 C_v),$$

$$\therefore C_v = -\frac{51}{29}.$$

Thus our diagnostic index is

$$S = x_u - \frac{51}{29}x_v.$$

For the two populations its mean values are

$$I_a = 13 - \frac{10(51)}{29} \simeq -4.6,$$

$$I_b = 15 - \frac{4(51)}{29} \simeq +8.0.$$

\*       \*       \*       \*

*Generalisation of the Evaluation.* In circumstances which call for the use of such a diagnostic index (*discriminant function*), we shall commonly have more than two measurements to combine, and we may write more generally for  $p$  of them

$$S = \sum_{n=1}^{n=p} C_n \cdot x_{j \cdot n}; \quad D = \sum_{n=1}^{n=p} C_n \cdot d_{j \cdot n} \quad . \quad . \quad . \quad . \quad (viii)$$

$$\frac{\partial V}{\partial C_k} = 2 \sum_{n=1}^{n=p} C_n \cdot m_{kn}.$$

Thus we have a set of equations of the form

$$M_{d \cdot 1} = \frac{(D - M_d)}{V} (C_1 \cdot m_{11} + C_2 \cdot m_{12} + C_3 \cdot m_{13} \dots C_p \cdot m_{1p}),$$

$$M_{d \cdot 2} = \frac{(D - M_d)}{V} (C_1 \cdot m_{21} + C_2 \cdot m_{22} + C_3 \cdot m_{23} \dots C_p \cdot m_{2p})$$

etc.

etc.

As before,  $m_{nn}$  is the sum of the two variances of the  $n$ th measurement,  $m_{nk}$  being the sum of the two covariances of the  $n$ th and  $k$ th. We are at liberty to solve by setting  $C_1 = 1$ , the constant  $(D - M_d) \div V$  being then irrelevant, and our equations thus take the form

$$M_{d \cdot n} = m_{n1} + C_2 \cdot m_{n2} + C_3 \cdot m_{n3}, \text{ etc.} \quad . \quad . \quad . \quad . \quad (ix)$$

The foregoing treatment sets forth how we initially determine the values of the weights we employ to get the best diagnostic index embodying different sets of measurements. Having done so, we may use it as a classificatory device. Thus we determine  $S$  in (viii) for a doubtful specimen and assign the latter to class A *if* the numerical value obtained lies nearer to  $I_a$ , its A-class mean value.

*Addendum.* Two of my younger colleagues who kindly read through the proofs of this chapter have expressed the misgiving that I may have overstressed the limitations of the Gaussian Theory of Errors as an instrument of research in biology and the social sciences. It is refreshing to recall the preface and argument of a still standard exposition of so-called regression. Therein Brunt (1917) states (*The Combination of Observations*):

The proof of the Normal Error Law has been based on Hagen's hypotheses regarding errors of observation. In most of the problems of Astronomy, Geodetics, and Physics the errors of observation satisfy the hypotheses, and the application of least square methods is justified. But cases may arise in which particular care is necessary in applying these methods.



This is especially true of Biological problems. For organic variability is the resultant of a large number of contributory causes, some of which may have a definite tendency to act always in one direction. The effect of such a bias is to produce an unsymmetrical frequency distribution, and the application of ordinary least square methods is then meaningless. It is thus in no way justifiable to regard Least Squares as a magical instrument applicable to all problems.

I recommend Brunt's book to any biologist or sociologist who entertains a reasonable scepticism about the credentials of a theory of curve fitting based on the pioneer work of Gauss and Hagen in the thirties of the last century when so recently imported into a domain entirely foreign to their intentions. The first few chapters are well worth reading for another reason. On all sides, we now hear that science has relinquished the quest for absolute truth by embracing the doctrine that its laws are merely statistical. This is at best a half truth, unless we exclude all forms of taxonomical enquiry from the title to rank as science. Even so, it is profoundly misleading. *Statistical* is an epithet with at least five different meanings in current educated speech. In the context of the assertion cited, it covers: (a) a calculus of aggregates (*e.g.* the kinetic theory of gases or the genetical theory of populations); (b) the Gaussian calculus of errors of observation; (c) a calculus of judgments. In this chapter we have seen reason for the doubt Brunt expresses and the need to re-examine the assumption that (b) and (c) have anything in common other than the algebraic devices they invoke. In Chapter 20 we shall see how little agreement exists w.r.t. assumptions common to (a) and (c).

What is equally relevant to the current claim stated above is that the Gaussian theory (*vide* Brunt, p. 34, l. 4) presupposes the existence of a *true* value as a foothold for any meaningful definition of *error* as such. Within the framework of its assumptions this true value is the *arithmetic mean* of an infinite number of trials. It is important to realise that this is the only consideration relevant to a justifiable identification of the mean with the *expected* value. The interchangeability of the terms in current statistical writing (including this book) is misleading in any other context. Outside the Gaussian domain, the mean—like the variance—can claim no special semantic status in preference to other parameters more or less usefully invoked in the formulation of sampling distributions.

## CHAPTER 18

# ELEMENTS OF ANALYSIS OF COVARIANCE AND OF FACTOR ANALYSIS

### 18.01 REGRESSION AS A STANDARDISING DEVICE

IN Chapter 17 we have seen that the least squares estimate of the constants of a linear law has a long history in the so-called exact sciences. On that account the concept of a physical law has cast—and still casts—a long shadow over the statistical theory of regression as applied to biological and sociological enquiries. From the viewpoint of the physicist, two issues are of paramount concern: (a) are the data of an experiment consistent with the coexistence of unavoidable experimental errors and of a law suggested by the data themselves or (and more often) by a particular hypothesis from a cognate domain of enquiry? (b) if so, what are the most reliable estimates of the definitive parameters, e.g. an elastic modulus or the E.M.F. of a standard cell?

The second question has in fact little meaning unless we predicate what is implicit in the statement of the first, i.e. that the major source of variation arises from random error of observation, instrumental or personal. As we have seen, this is rarely, if ever, true of situations which arise in sociological enquiry; and it is by no means always true in the domain of experimental biology. If we plot sociological and biological data in conformity with the traditional technique of least squares, we rarely do so to prescribe a figure comparable to a physical constant. We do so to decide whether some putative causal agency exerts a real influence or merely whether there is some causal *nexus* responsible for concomitant variation of different score sets.

The student will experience little difficulty in appreciating this shift of interest, if we here digress to discuss a typical situation in which the biologist may invoke the technique of regression with more or less advantage. We shall suppose that we are investigating the response of 2 groups of animals on a different diet to one and the same drug. We have then to take stock of the fact that individuals of different size will not respond equally to the same dosage of the drug. In the absence of any diet effect, the administration of the same dose to each individual might therefore result in a group mean difference, since it would very rarely happen that the mean weights of the groups would be identical. The investigator can sidestep this pitfall in several ways:

- (a) by choosing animals of so nearly the same body weight that any such source of variation would be trivial;
- (b) by pairing off individuals of nearly the same body weight in each group and by giving each pair the same dosage;
- (c) by pre-adjustment of individual dosage based on previous knowledge concerning the relationship of dosage itself to body weight for a response of fixed magnitude;
- (d) by using information gained in the course of the experiment to adjust the figures accordingly.

The first is the ideal of the worker at home with his materials; but is sometimes impracticable. Some combination of (b) and (c) is then the best course to pursue; and (d), which is a *pis aller* in laboratory enquiry, raises issues we shall explore more fully in connection with the technique known as *Analysis of Covariance*. Essentially, the latter is a battery of significance





Example 2. For the following figures regression is roughly linear:

	$w$	$d$	$dw$	$w^2$	Observed $d$	Calculated $d$
	3	9	27	9	9	10
	3	10	30	9	10	10
	4	11	44	16	11	10.7
	5	13	65	25	13	11.4
	7	12	84	49	12	12.75
Total	22	55	250	108		

$$\text{Cov}(d, w) = \frac{8}{5}; \quad V_w = \frac{56}{25}; \quad k_{dw} = \frac{5}{7}; \quad C = \frac{5.5}{7} = 7.86;$$

$$d_r = \frac{5}{7}(w + 11).$$

Having conducted a pilot enquiry involving  $p$  paired observations to evaluate the constants in (ii) above, and this will, of course, commit us to a test of linearity, we can state the approximate margin of error entailed in assigning the correct value to  $d$  by recourse to the formula of (xxi) in 17.04, viz.:

$$s_e^2 \left[ 1 + \frac{1}{p} + \frac{(w - M_w)^2}{p \cdot V_w} \right] \quad \text{. . . . .} \quad (\text{iii})$$

If we denote the actual dosages for a given body weight in the standard graph of our pilot experiment by  $d_{ij}$ , we define  $s_e^2$  for  $p$  paired values in all and  $c$  different values of  $w$  in accordance with (xv) of 17.03 by the equation

$$s_e^2 = \frac{p}{p-2} \sum_{j=1}^p \sum_{i=1}^c (d_{ij} - d_{r.i})^2 \quad \text{. . . . .} \quad (\text{iv})$$

By recourse to (iii) and (iv) we can assess the legitimate confidence with which we may adjust dosage to body weight *in advance* in accordance with (c) above. The end in view is characteristic of numerous situations in which it is our concern to eliminate a known source of variation otherwise likely to vitiate a judicious evaluation of another; but the task does not always admit of disposal by recourse to results obtained once and for all by a pilot enquiry of the sort illustrated above. In some circumstances the existence of an additional contributory source of variation may not be apparent beforehand. It may then be possible to be wise after the event. If we do suspect the existence of a second source of variability irrelevant to our end in view, we may design our experiment to provide us simultaneously with sufficient information about its contribution to justify a confident answer to the main question.

When our concern is indeed to evaluate 2 or more sources of variation within the framework of one and the same set of observations, it may happen that each (e.g. dose and size or percentage literacy and family income) is expressible quantitatively by reference to a system of scores, and that each is conceivably contributory to a quantifiable response (e.g. hours of sleep or infant mortality). If so, we may separately assess their effects by the method of multiple regression (*vide* 17.07) or, in a less satisfactory way, by recourse to partial correlation (*vide* 9.04 in Vol. I). In the symbols of an earlier paragraph (p. 765) our empirical regression





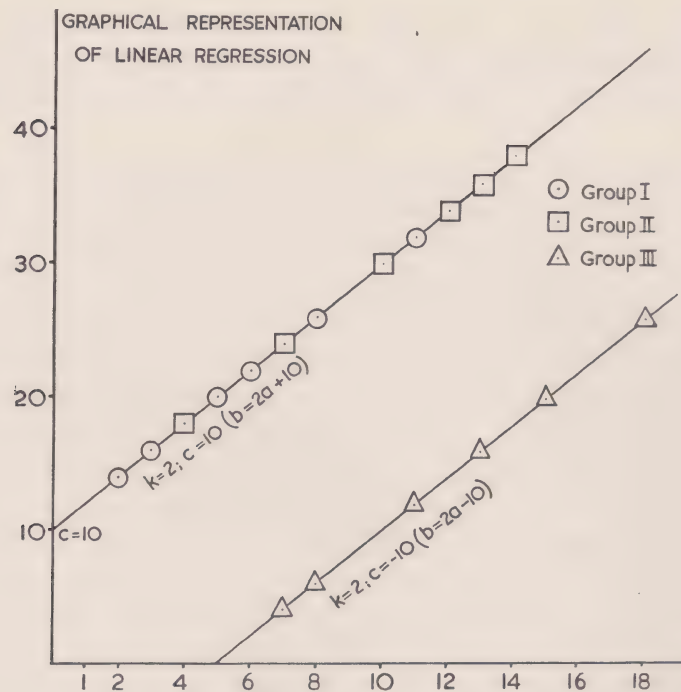


FIG. 123. Linear Regression.

This figure refers to three sets of paired scores exhibiting an exactly linear relationship.

Total	I				II				III			
	<i>a</i>	<i>b</i>	<i>ab</i>	<i>a</i> <sup>2</sup>	<i>a</i>	<i>b</i>	<i>ab</i>	<i>a</i> <sup>2</sup>	<i>a</i>	<i>b</i>	<i>ab</i>	<i>a</i> <sup>2</sup>
	2	14	28	4	4	18	72	16	7	4	28	49
	3	16	48	9	7	24	168	49	8	6	48	64
	5	20	100	25	10	30	300	100	11	12	132	121
	6	22	132	36	12	34	408	144	13	16	208	169
	8	26	208	64	13	36	468	169	15	20	300	225
	11	32	352	121	14	38	532	196	18	26	468	324
	35	130	868	259	60	180	1948	674	72	84	1184	952

The determination of the regression lines from the appropriate statistical parameters is as follows :

$$\text{Group I. } V_a = \frac{6(259) - (35)^2}{36} = \frac{329}{36}; \text{Cov}(a, b) = \frac{6(868) - (35)(130)}{36} = \frac{658}{36}; k_{ba} = \frac{658}{329} = 2;$$

$$C = M_b - k_{ba}M_a = \frac{130}{6} - \frac{2(35)}{6} = 10.$$

$$\text{Group II. } V_a = \frac{6(674) - (60)^2}{36} = \frac{444}{36}; \text{Cov}(a, b) = \frac{6(1948) - (60)(180)}{36} = \frac{888}{36}; k_{ba} = \frac{888}{444} = 2;$$

$$C = M_b - k_{ba}M_a = \frac{180}{6} - \frac{2(60)}{6} = 10.$$

$$\text{Group III. } V_a = \frac{6(952) - (72)^2}{36} = \frac{528}{36}; \text{Cov}(ab) = \frac{6(1184) - (72)(84)}{36} = \frac{1056}{36}; k_{ba} = \frac{1056}{528} = 2;$$

$$C = M_b - k_{ba}M_a = \frac{84}{6} - \frac{2(72)}{6} = -10.$$



We may denote the corresponding means of the two groups by  $M_{a.1}$ ,  $M_{a.2}$  and  $M_{b.1}$ ,  $M_{b.2}$ . By hypothesis,  $k_{ba.1} = k_{ba} = k_{ba.2}$ , i.e. the regression coefficients are the same for each group. Hence we have

$$M_{b.1} - k_{ba} \cdot M_{a.1} = C = M_{b.2} - k_{ba} \cdot M_{a.2} \quad . \quad . \quad . \quad (vi)$$

Within the framework of the foregoing assumptions, i.e. that regression is linear and that the regression coefficients are identical, the two  $B$ -score means defined by (vi) will therefore be equal if, and only if,  $M_{a.1} = M_{a.2}$ . Thus they will be the same if each group experiences the same food mean intake ( $M_{sa}$ ) as the pooled assemblage of both groups. We may therefore define standardised (or *adjusted*) means ( $M_{sb.1}$  and  $M_{sb.2}$ ) by the relations

$$M_{sb.1} = k_{ba} \cdot M_{sa} + C = M_{sb.2}.$$

From (vi) above

$$M_{sb.1} = M_{b.1} + k_{ba}(M_{sa} - M_{a.1}) \quad . \quad . \quad . \quad (vii)$$

Similarly

$$M_{sb.2} = M_{b.2} + k_{ba}(M_{sa} - M_{a.2}) \quad . \quad . \quad . \quad (viii)$$

The standardised group mean growth score is therefore obtainable by *adding to the crude group mean the product of the group regression coefficient and the difference between the food-intake grand mean of the pooled sample and the group food-intake mean*.

Let us now suppose that diet II does have a specific additive effect in addition to its non-specific action on appetite as shown by the fact that  $M_{sb.2} > M_{sb.1}$ . If we denote as  $F_2$  the growth increment due to this specific food factor, our equations of score components become

$$\begin{aligned} x_{b.1} &= k_{ba} \cdot x_{a.1} + C \quad \text{and} \quad x_{b.2} = k_{ba} \cdot x_{a.2} + C + F_2, \\ \therefore M_{sb.2} &= k_{ba} \cdot M_{sa} + C + F_2 = M_{sb.1} + F_2 \quad . \quad . \quad . \quad (ix) \end{aligned}$$

If diet II has a specific effect, the standardised group growth mean will therefore be greater than that of group I. A numerical example will assist to clarify the foregoing arguments.

\* \* \* \* \*

*Numerical Example.* Table 1 shows 3 series of  $p$  ( $= 4$ ) paired scores, regression being exactly linear for each series. The reader may check as an exercise the values given for the regression equations ( $b = 3a + 5$ ) which are identical for the first two series. The *slope* ( $k_{ba} = 3$ ) for the third is the same

TABLE 1

	Series I				Series II				Series III			
	<i>a</i>	<i>b</i>	<i>ab</i>	<i>a</i> <sup>2</sup>	<i>a</i>	<i>b</i>	<i>ab</i>	<i>a</i> <sup>2</sup>	<i>a</i>	<i>b</i>	<i>ab</i>	<i>a</i> <sup>2</sup>
	4	17	68	16	8	29	232	64	9	37	333	81
	5	20	100	25	9	32	288	81	10	40	400	100
	7	26	182	49	10	35	350	100	12	46	552	144
	10	35	350	100	12	41	492	144	15	55	825	252
Totals	26	98	700	190	39	137	1362	389	46	178	2110	550
	$M_b = 24.5; k_{ba} = 3$ $M_a = 6.5; C = 5$ $b = 3a + 5$				$M_b = 34.25; k_{ba} = 3$ $M_a = 9.75; C = 5$ $b = 3a + 5$				$M_b = 44.5; k_{ba} = 3$ $M_a = 11.5; C = 10$ $b = 3a + 10$			

I and II pooled ( $p = 8$ )

	<i>a</i>	<i>b</i>	<i>ab</i>	<i>a</i> <sup>2</sup>
	26	98	700	190
	39	137	1362	389
	..	...	....	...
Total	65	235	2062	579

$$M_b = 29.375; k_{ba} = 3;$$

$$M_a = 8.125; C = 5; b = 3a + 5.$$

I, II, III pooled ( $p = 12$ )

	<i>a</i>	<i>b</i>	<i>ab</i>	<i>a</i> <sup>2</sup>
	26	98	700	190
	39	137	1362	389
	46	178	2110	550
Total	111	413	4172	1129

$$M_b = 34.42; (k_{ba} \simeq 3.44);$$

$$M_a = 9.25; (C \simeq 2.6) \dots$$

as for the other two, but the origin ( $C = 10$ ) of the distribution is different. This is equivalent to adding a specific factor  $F_3 = 5$  to each  $B$ -score computed from the regression equation of the other 2 series. If we pool I and II, we obtain the pooled mean  $M_a = 8.125$ , whence we arrive at the following result (Table 2):

TABLE 2

	$M_b$ crude	$M_b$ standardised
I	24.5	$24.5 - 3(6.5 - 8.125) = 29.375$
II	34.25	$34.5 - 3(9.75 - 8.125) = 29.375$

If we now standardise the scores by reference to the value of  $M_a (= 9.25)$  for the entire pool of data, we have

	$M_b$ crude	$M_b$ standardised
I	24.5	$24.5 - 3(6.5 - 9.25) = 32.75$
II	34.25	$34.25 - 3(9.75 - 9.25) = 32.75$
III	44.5	$44.5 - 3(11.5 - 9.25) = 37.75$

\* \* \* \* \*



Thus the standardised  $B$ -scores for the first 2 series of our numerical example are identical whether we standardise them for comparison with one another alone or for comparison with the third. The standardised  $A$ -score for III exceeds them by the series factor ( $F_3 = 5$ ). This must be so, as we see if we write the equations

$$\begin{aligned} b_1 &= k_1 a_1 + C &= 3a_1 + 5 + 0; & M_{b.1} = 3M_{a.1} + 5; \\ b_2 &= k_2 a_2 + C &= 3a_2 + 5 + 0; & M_{b.2} = 3M_{a.2} + 5; \\ b_3 &= k_3 a_3 + C + F_3 &= 3a_2 + 5 + 5; & M_{b.3} = 3M_{a.3} + 10. \end{aligned}$$

Whence the standardised mean  $B$ -scores are

$$\begin{aligned} \text{I. } M_{b.1} - 3(M_{a.1} - M_a) &= 3M_{a.1} + 5 - 3(M_{a.1} - M_a) = 3M_a + 5. \\ \text{II. } M_{b.2} - 3(M_{a.2} - M_a) &= 3M_{a.2} + 5 - 3(M_{a.2} - M_a) = 3M_a + 5. \\ \text{III. } M_{b.3} - 3(M_{a.3} - M_a) &= 3M_{a.3} + 10 - 3(M_{a.3} - M_a) = 3M_a + 10. \end{aligned}$$

In terms of assessment of treatment we may sum up the foregoing remarks about standardisation as follows. We suppose that we have before us, for each of several treated groups, paired values of the responses ( $b$ ) of different individuals and of some correlated score ( $a$ ). Our  $A$ -score means vary from group to group and our aim is to assess how far this circumstance suffices to account for the treatment group mean differences w.r.t. the response itself. *In the absence of any residual source of variation*, we may say that

- (i) group means adjusted by the method described above will be identical, if treatment *per se* has no effect;
- (ii) if treatment exerts an independent specific effect, being then such as to shift the origin of the regression from  $C$  to  $C + F$ , its influence will appear as an increment (or decrement, if  $F$  is negative) numerically equivalent to  $F$ .

In biological and sociological enquiry, it is, of course, impossible to exclude residual sources of variation affecting the responses of individuals or communities; but we can sometimes justifiably assume that their collective effect is random in the sense that positive and negative deviations from the regression mean resulting therefrom cancel out in the long run. In practice, therefore, standardising our data by recourse to the regression equation is unlikely to yield adjusted means which are exactly equal when treatment has no effect. What it can do is to get the meaning of the crude data into sharper focus. If the effect of standardisation is to reduce the group mean differences very noticeably, we have reason to suspect that the residual differences are attributable to random residual variation, being *insignificant* in that sense. Having removed the effect of the uncontrolled variable  $A$ , we have thus to ascertain whether the residual variation is still accountable without invoking the assumption that treatment is efficacious. This is the major objective of the statistical technique known as analysis of covariance.

## 18.02 ANALYSIS OF COVARIANCE

The need for a technique such as analysis of covariance arises in circumstances when :

- (a) we wish to determine whether some qualitative criterion of classification, e.g. treatment, significantly contributes to the variation of a score  $B$ , e.g. gain of body weight;
- (b) we also have reason to believe that the score  $B$  depends in part at least on another variable  $A$ , e.g. food intake, which is not under direct control and is therefore inconstant w.r.t. groups distinguished by the criterion of classification and unlikely to have the same mean value in any two of them.

When it is not indeed possible to eliminate the effect of such variation a true effect associated with the criterion of classification signifies that the relation between  $A$  and  $B$  in different groups is not the same. If we can score  $A$  so that regression of  $B$  on  $A$  is linear, a true difference may show up in either or both of two ways :

- (i) regression is not uniform, i.e. the regression coefficients are not all identical ;
  - (ii) there is a specific and group effect, i.e. regression lines do not have the same origin.
- For either or both reasons the adjusted means will in general be different.

To assess the significance of adjusted means we thus need two different tests which bear directly on the issue raised above ; but the performance of either presupposes that we can safely assume regression within the groups to be linear. This would raise no new issue, if we were free to pick and choose our  $A$ -score values, as we can do when they are amenable to direct control ; but if so, we could design our enquiry without raising the problem we now face. Otherwise, the test for linearity based on (xiv) of 17.04 may fail us, because the  $F$ -ratio is indeterminate when each different  $B$ -score value within a group goes with a different value of the  $A$ -score.

In laboratory practice, we shall rarely be concerned with comparison of more than 2 treatment procedures at once, but in certain types of trials it may be advantageous to deal with more than two groups of paired scores. We shall therefore regard the comparison of 2 groups as a particular case of a more general pattern. We may visualise the lay-out (Table 1) for 3 groups as below :

TABLE 1

Group	I		II		III	
	$A$	$B$	$A$	$B$	$A$	$B$
	$a_{1.1}$	$b_{1.1}$	$a_{1.2}$	$b_{1.2}$	$a_{1.3}$	$b_{1.3}$
	$a_{2.1}$	$b_{2.1}$	$a_{2.2}$	$b_{2.2}$	$a_{2.3}$	$b_{2.3}$
	$a_{3.1}$	$b_{3.1}$	$a_{3.2}$	$b_{3.2}$	$a_{3.3}$	$b_{3.3}$
	....	....	$a_{4.2}$	$b_{4.2}$	$a_{4.3}$	$b_{4.3}$
	....	....	$a_{5.2}$	$b_{5.2}$	....	....
Means	$M_{a.1}$	$M_{b.1}$	$M_{a.2}$	$M_{b.2}$	$M_{a.3}$	$M_{b.3}$
No. of paired scores	$p_1 = 3$		$p_2 = 5$		$p_3 = 4$	

The essentially new question such a table prompts us to ask is whether there is a group effect. If so, we may also ask, is this because regression is not uniform or because the group effect is additive if regression is indeed uniform as defined above ? It may also be useful in certain circumstances to refute the suspicion that variation w.r.t. the  $A$ -score *per se* contributes anything appreciably to variation w.r.t. the  $B$ -score. This calls for the addition of another test to the battery. The entire battery of appropriate significance tests is a sequence in which the answer obtained from one decides whether it is worth while to ask the next. We may list them in this order :



- (i) if regression is linear is there a group effect of either sort ?
- (ii) if so, is regression uniform from group to group ?
- (iii) if so, is there an additive group effect ?
- (iv) is there true within-group regression ?

The logical order of procedure is a little puzzling to the beginner to whose difficulties the practice of exhibiting it against a background of elaborate computations adds needlessly. The necessary computations for the tests are very laborious, and the arithmetical order of procedure involves short cuts which have nothing to do with logical precedence. It is permissible to wonder how any student first confronted with paradigms chosen from agricultural trials or the like can hope to emerge from such a maze with any clear conception of the framework of assumptions relevant to correct application of the technique.

As in the foregoing examination of significance tests for regression estimates, the procedure prescribed is :

- (i) to formulate independent estimates of the true variance of the putative common universe of *e*-score (*residual*) components with a view to the use of the *F*-test in accordance with principles by now familiar ;
- (ii) to employ as the denominator of such a variance ratio (*F*) a *yardstick* statistic which necessarily depends on residual variation ( $\sigma_e^2$ ) alone ;
- (iii) to employ as the numerator of the *F*-ratio a statistic whose expected value will exceed  $\sigma_e^2$  if the null hypothesis is false.

In what follows we proceed in the same way with this qualification. It may happen that we can formulate an independent statistic whose expected value will be *less* than that of the yardstick statistic if the null hypothesis is false. To use the *F*-table intelligently we must then employ the former as the denominator of the *F*-ratio and the latter as the numerator.

*Notation.* In defining appropriate expressions for the numerator or denominator of the *F*-ratio, we have had to assume that we are sampling in accordance with the principle of the fixed-*A* set. In this context, the principle presupposes a doubly stratified universe, since we have to assume that the *A*-score distribution is fixed for each set of paired scores as in the sample. If this is clear we may drop the subscripts *c* and *s* except when we need to distinguish the true value ( $k_{ba}$ ) from the sample value ( $k_{ba \cdot cs}$ ) of the regression coefficient. Our code will be as follows :

	Pooled Sample	Within the <i>k</i> th sub-sample
No. of paired scores . . . . .	$p$	$p_k$
Mean <i>A</i> -score . . . . .	$M_a$	$M_{a \cdot k}$
<i>A</i> -score Variance . . . . .	$V_a$	$V_{a \cdot k}$
Mean <i>B</i> -score . . . . .	$M_b$	$M_{b \cdot k}$
<i>B</i> -score Variance . . . . .	$V_b$	$V_{b \cdot k}$

For *h* sets of paired scores we may designate the operation of extracting a mean value as

$$\frac{1}{p} \sum_{k=1}^{k=h} p_k(\dots) \equiv E_h(\dots),$$

$$\therefore p \cdot E_h\left(\frac{p_k - c}{p_k}\right) = p - ch \quad \text{and} \quad p \cdot E_h\left(\frac{1}{p_k}\right) = h \quad . \quad . \quad . \quad (i)$$

For the operation of extracting the within-set mean of all  $p_k$  values we may likewise write

$$\sum_{u=1}^{u=p_k} ( \dots ) \equiv E_w( \dots ).$$

For the expected value of a parameter  $W$  within the fixed- $A$  set as a whole we shall use  $E_c(W)$ , so that the expected value of the mean  $e$ -score variance within the set is

$$\begin{aligned} E_c \cdot M(V_{e \cdot h}) &= E_c \cdot E_h(V_{e \cdot h}) = E_h \cdot E_c(V_{e \cdot h}), \\ \therefore E_c \cdot M(V_{e \cdot h}) &= E_h \left( \frac{p_h - 1}{p_h} \sigma_e^2 \right) = \frac{p - h}{p} \sigma_e^2 \quad \dots \quad (ii) \end{aligned}$$

To complete our code we must make explicit the putative components of the  $B$ -scores. If there is uniform regression and no group effect associated with our qualitative criterion of classification, i.e. all sets of paired scores come from the same bivariate universe, we may write

$$b = e + F_a + C.$$

If there is a group effect either  $F_a$  or  $C$  varies from group to group, indeed both may do so; and we may distinguish

$$\begin{aligned} b &= e + F_a + C_h && \text{regression uniform, additive group effect present;} \\ b &= e + F_{a \cdot h} + C && \text{regression coefficient variable from group to group, no other group effect;} \\ b &= e + F_{a \cdot h} + C_h && \text{regression coefficient variable and additive group factor present.} \end{aligned}$$

If regression is also linear we may write  $F_a = k_{ba} \cdot a$  or  $F_{a \cdot h} = k_{ba \cdot h} \cdot a$  in the above, as the case may be. The accompanying table of  $B$ -score components (Table 2) fills in any essential gaps. By reference thereto we can at once derive a result which will clarify subsequent reasoning. When  $h$  sets of paired scores come from the same bivariate universe, we have before us  $h$  paired mean  $A$ -scores and mean  $B$ -scores; and we may define in the usual way a coefficient of regression of the mean  $B$ -score on the mean  $A$ -score. The expected value of this coefficient ( $k_{m \cdot c}$ ) is the true regression coefficient ( $k_{ba}$ ). This is deducible from the following considerations. If regression is linear and uniform, in the absence of a group effect

$$\begin{aligned} M_{b \cdot h} &= M_{e \cdot h} + k_{ba} \cdot M_{a \cdot h} + C; \\ \text{Cov}(M_a, M_b) &= E_h(M_{a \cdot h} - M_a)M_{b \cdot h} \\ &= E_h(M_{a \cdot h} - M_a)M_{e \cdot h} + k_{ba} \cdot E_h(M_{a \cdot h} - M_a)M_{a \cdot h}. \end{aligned}$$

In this expression

$$\begin{aligned} E_h(M_{a \cdot h} - M_a)M_{a \cdot h} &= E_h(M_{a \cdot h}^2) - M_a E_h(M_{a \cdot h}) = E_h(M_{a \cdot h}^2) - M_a^2 = V(M_{a \cdot h}); \\ \text{Cov}(M_a, M_b) &= E_h(M_{a \cdot h} - M_a)M_{e \cdot h} + k_{ba} \cdot V(M_{a \cdot h}). \end{aligned}$$

The expected value of the first term on the right, being the covariance of the mean  $A$ -scores and the mean of the independent residual  $e$ -scores, is zero and within the fixed- $A$  set:

$$E_c \cdot \text{Cov}(M_a, M_b) = k_{ba} \cdot V(M_{a \cdot h}).$$

If  $k_{m \cdot c}$  is the observed *sample* value of the regression coefficient of the mean  $B$ -score on the mean  $A$ -score:

$$\begin{aligned} k_{m \cdot c} &= \frac{\text{Cov}(M_a, M_b)}{V(M_{a \cdot h})}, \\ \therefore E_c(k_{m \cdot c}) &= k_{ba} \quad \dots \quad (iii) \end{aligned}$$



TABLE 2

	<i>Individual Score</i> (b)	<i>Set Mean</i> ( $M_{b..h}$ )	<i>Within-set Variance</i> ( $V_{b..h}$ )	<i>Variance of Means</i> $V(M_{b..h})$	<i>Total Variance</i> $V_b$
<i>No restriction</i>	$e + F_{a..h} + C_h$	$M_{e..h} + M_{f..h} + M_{c..h}$	$V_{e..h} + V_{f..h}$	$V(M_{e..h}) + V(M_{f..h}) + V(M_{c..h})$	$V_e + V_f + V_c$
<i>Regression uniform</i>	$e + F_a + C_h$	$M_{e..h} + M_f + M_{c..h}$	$V_{e..h} + V_f$	$V(M_{e..h}) + V(M_{c..h})$	$V_e + V_c$
<i>Ditto, no additive group effect</i>	$e + F_a + C$	$M_{e..h} + M_f + C$	$V_{e..h} + V_f$	$V(M_{e..h})$	$V_e$
<i>Regression linear</i>	$e + k_{ba..h}a + C_h$	$M_{e..h} + k_{ba..h}M_{a..h} + C_h$	$V_{e..h} + k_{ba..h}^2 V_{a..h}$	$V(M_{e..h}) + V(k_{ba..h}M_{a..h}) + V(M_{c..h})$	$V_e + V(k_{ba..h}^2 a) + V_c$
<i>Regression linear and uniform</i>	$e + k_{ba..a} + C_h$	$M_{e..h} + k_{ba..h}M_a + C_h$	$V_{e..h} + k_{ba..h}^2 V_{a..h}$	$V(M_{e..h}) + k_{ba..h}^2 V(M_a) + V_c$	$V_e + k_{ba..h}^2 V_a + V_c$
<i>Ditto, no additive group effect</i>	$e + k_{ba..a} + C$	$M_{e..h} + k_{ba..h}M_a + C$	$V_{e..h} + k_{ba..h}^2 V_{a..h}$	$V(M_{e..h}) + k_{ba..h}^2 V(M_a)$	$V_e + k_{ba..h}^2 V_a$

$$E_c(s_3^2) = \sigma_\epsilon^2; \quad s_3^2 = \frac{S_3}{p-2h}; \quad S_3 = M(1-r_{ab,h}^2)p \cdot V_{b,h} \quad . \quad . \quad . \quad (\text{ix})$$



The expected value of the statistic ( $s_3^2$ ) depends *neither* on the assumption of uniform regression ( $k_{ba \cdot h} = k_{ba}$ ) *nor* on the assumption that there is no group effect ( $C_h = C$ ). If regression is uniform, we may indeed obtain a fourth statistic whose value does not depend on the existence of a group effect by using the square ( $r_{ab \cdot m}^2$ ) of the mean within-group correlation coefficient for the mean square  $M(r_{ab \cdot h}^2)$  of the *within*-group correlation coefficient in (vii). We define  $r_{ab \cdot m}^2$  as follows :

$$r_{ab \cdot m}^2 = \frac{[M \text{Cov}(a, b)]^2}{M(V_{a \cdot h})M(V_{b \cdot h})},$$

$$\therefore r_{ab \cdot m}^2 \cdot M(V_{b \cdot h}) = \frac{[M \text{Cov}(a, b)]^2}{M(V_{a \cdot h})} = M(r_{ab \cdot m}^2 \cdot V_{b \cdot h}) \quad \dots \quad (x)$$

Thus the statistic whose expected value we shall now determine is

$$M(1 - r_{ab \cdot m}^2)V_{b \cdot h} = \frac{S_4}{p} = M(V_{b \cdot h}) - r_{ab \cdot m}^2 \cdot M(V_{b \cdot h}) \quad \dots \quad (xi)$$

In this expression uniform linear regression implies

$$M(V_{b \cdot h}) = M(V_{e \cdot h}) + k_{ba}^2 \cdot M(V_{a \cdot h}).$$

Whence within the fixed- $A$  set from (ii)

$$E_c \cdot M(V_{b \cdot h}) = \frac{p-h}{p} \sigma_e^2 + k_{ba}^2 \cdot M(V_{a \cdot h}) \quad \dots \quad (xii)$$

In (x) above,

$$M \text{Cov}(a, b) = M \text{Cov}(a, e) + k_{ba} \cdot M(V_{a \cdot h}),$$

$$\therefore r_{ab \cdot m}^2 \cdot M(V_{b \cdot h}) = \frac{[M \text{Cov}(a, e)]^2}{M(V_{a \cdot h})} + 2k_{ba} \cdot M \text{Cov}(a, e) + k_{ba}^2 \cdot M(V_{a \cdot h}).$$

In this expression the last term, the coefficient of the covariance in the second term and the denominator of the first are constants of the fixed- $A$  set. Since the expected value of  $\text{Cov}(a, e)$  is zero, we may therefore write

$$E_c \cdot r_{ab \cdot m}^2 \cdot M(V_{b \cdot h}) = \frac{E_c[M \text{Cov}(a, e)]^2}{M(V_{a \cdot h})} + k_{ba}^2 \cdot M(V_{a \cdot h}).$$

Whence from (xii), if regression is uniform,

$$E_c(1 - r_{ab \cdot m}^2) \cdot M(V_{b \cdot h}) = \frac{p-h}{p} \sigma_e^2 - \frac{E_c[M \text{Cov}(a, e)]^2}{M(V_{a \cdot h})} \quad \dots \quad (xiii)$$

To evaluate the second term on the right it will be convenient to put  $M \cdot \text{Cov}(a, e) = Q$ , so that  $E_c(Q) = 0$ . By definition of variance, we may therefore write  $E_c(Q^2) = \sigma_Q^2$ . If  $z_m = \text{Cov}(a, e)$  within the  $m$ th set :

$$M \text{Cov}(a, e) = Q = \sum_{m=1}^{m=h} \frac{p_m}{p} z_m.$$

Each component  $z_m$  of  $Q$  is independent of any other, being referable to different sub-universes. Hence if  $\sigma_{z \cdot m}^2$  is the variance of the distribution of  $z_m$ ,

$$E_c[M \text{Cov}(a, e)]^2 = \sigma_Q^2 = \sum_{m=1}^{m=h} \frac{p_m^2}{p^2} \sigma_{z \cdot m}^2 \quad \dots \quad (xiv)$$

TABLE 3

*Key for Computation (Sums of Squares and Products).*

$S_{aa \cdot w} = \sum_{u=1}^{u=p_w} (a_{uw} - M_{a \cdot w})^2$	$S_{bb \cdot w} = \sum_{u=1}^{u=p_w} (b_{uw} - M_{b \cdot w})^2$	$S_{ba \cdot w} = \sum_{u=1}^{u=p_w} (a_{uw} - M_{a \cdot w})(b_{uw} - M_{b \cdot w})$
$S_{aa \cdot m} = \sum_{w=1}^{w=h} p_w (M_{a \cdot w} - M_a)^2$	$S_{bb \cdot m} = \sum_{w=1}^{w=h} p_w (M_{b \cdot w} - M_b)^2$	$S_{ba \cdot m} = \sum_{w=1}^{w=h} p_w (M_{a \cdot w} - M_a)(M_{b \cdot w} - M_b)$
$S_{aa \cdot o} = \sum_{w=1}^{w=h} \sum_{u=1}^{u=p_w} (a_{uw} - M_a)^2$	$S_{bb \cdot o} = \sum_{w=1}^{w=h} \sum_{u=1}^{u=p_w} (b_{uw} - M_b)^2$	$S_{ba \cdot o} = \sum_{w=1}^{w=h} \sum_{u=1}^{u=p_w} (a_{uw} - M_a)(b_{uw} - M_b)$
$S_{aa \cdot p} = \sum_{w=1}^{w=h} S_{aa \cdot w}$	$S_{bb \cdot p} = \sum_{w=1}^{w=h} S_{bb \cdot w}$	$S_{ba \cdot p} = \sum_{w=1}^{w=h} S_{ba \cdot w}$
$S_1 = S_{bb \cdot o} - \frac{S_{ab \cdot o}^2}{S_{aa \cdot o}};$	$S_2 = S_{bb \cdot m} - \frac{S_{ba \cdot m}^2}{S_{aa \cdot m}};$	
$S_3 = \sum_{w=1}^{w=h} \left( S_{bb \cdot w} - \frac{S_{ba \cdot w}^2}{S_{aa \cdot w}} \right);$	$S_4 = S_{bb \cdot p} - \frac{S_{ba \cdot p}^2}{S_{aa \cdot p}}.$	



In accordance with (v) and (x) of 17.03

$$z_m = (k_{ba \cdot cs} - k_{ba})V_{a \cdot m};$$

$$\sigma_{z \cdot m}^2 = \frac{\sigma_e^2}{p_m \cdot V_{a \cdot m}} (V_{a \cdot m}^2) = \frac{\sigma_e^2 \cdot V_{a \cdot m}}{p_m}.$$

Whence from (xiv)

$$\sigma_q^2 = \frac{\sigma_e^2}{p} \sum_{m=1}^{m=h} \frac{p_m}{p} V_{a \cdot m},$$

$$\therefore E_c [M \text{Cov}(a, e)]^2 = \frac{\sigma_e^2}{p} M(V_{a \cdot h}).$$

Thus from (xiii) above

$$E_c(1 - r_{ab \cdot m}^2) \cdot M(V_{b \cdot h}) = \frac{p - h - 1}{p} \sigma_e^2 \quad . \quad . \quad . \quad (xv)$$

We may thus define a fourth statistic which is an unbiased estimate of  $\sigma_e^2$  if regression is uniform and linear regardless of the presence or absence of a group factor:

$$E(s_4^2) = \sigma_e^2; \quad s_4^2 = \frac{S_4}{p - h - 1}; \quad S_4 = (1 - r_{ab \cdot m}^2)p \cdot M(V_{b \cdot h}) \quad . \quad . \quad (xvi)$$

We can combine  $S_1$  of (i),  $S_3$  of (ix) and  $S_4$  of (xvi) to obtain other statistics which are unbiased estimates of  $\sigma_e^2$  on the assumption that

(i) regression is linear

$$E(s_5^2) = \sigma_e^2; \quad s_5^2 = \frac{S_5}{2h - 2}; \quad S_5 = S_1 - S_3 \quad . \quad . \quad . \quad (xvii)$$

(ii) regression is linear and uniform

$$E(s_6^2) = \sigma_e^2; \quad s_6^2 = \frac{S_6}{h - 1}; \quad S_6 = S_4 - S_3 \quad . \quad . \quad . \quad (xviii)$$

(iii) regression is linear and uniform, no other group effect

$$E(s_7^2) = \sigma_e^2; \quad s_7^2 = \frac{S_7}{h - 1}; \quad S_7 = S_1 - S_4 \quad . \quad . \quad . \quad (xix)$$

In specifying the assumptions subject to which the several statistics defined above are unbiased estimates of  $\sigma_e^2$ , we have not indicated what is of pivotal importance if we wish to prescribe an  $F$ -test in accordance with the procedure outlined above. Of those defined by the foregoing equations,  $s_3^2$  being referable exclusively to variation within the group is an unbiased estimate of  $\sigma_e^2$  whether regression is uniform ( $k_{ba \cdot h} = k_{ba}$ ) or not and whether there is or is not ( $C_h = C$ ) a group effect involving a shift of origin of the score distribution. It is thus the fundamental yardstick statistic; but if we are content with the outcome of a test of uniformity based thereon, we may proceed to use  $s_4^2$  as a yardstick statistic. The effect of variation among the values of  $r_{ab \cdot h}$  will be to make the expected value of the square of the mean within-group regression coefficient ( $r_{ab \cdot m}$ ) *greater* than it would otherwise be. We may therefore write

$$\begin{aligned} \text{regression linear and uniform} \quad & E(s_4) = E(s_3); \quad E(s_6) = E(s_4). \\ \text{regression linear, not uniform} \quad & E(s_4) > E(s_3); \quad E(s_6) > E(s_4). \end{aligned}$$

The statistic  $s_1^2$  of (iv) being referable to  $(1 - r_{ab}^2)V_b$  will diminish if the set-up is such as to increase the expected value of the common  $r_{ab}$ . Either differences of the regressions *inter se* or an additive group factor will make the latter greater than otherwise, whence we can write

*regression linear and uniform, no additive effect:*  $E(s_5) = E(s_3); E(s_7) = E(s_4);$

*regression linear with one group effect or both:*  $E(s_5) < E(s_3);$

*regression linear and uniform with additive effect:*  $E(s_7) < E(s_4).$

This leaves us with  $s_2^2$  which depends on the regression of the paired mean scores. Like  $s_1^2$  its value depends on variability of both regression coefficients within groups and the presence or absence of an additive group effect. Either sort of group effect will diminish its value. We may therefore state

*regression linear and uniform without additive effect:*  $E(s_2) = E(s_4);$

*regression linear and uniform with additive effect:*  $E(s_2) < E(s_4).$

If we have reason to ask whether there would still be a significant correlation between  $A$  and  $B$  in the absence of a group effect we may confine our attention to the fraction of total variance which is not affected by either sort of variation which may arise from the group classification, *viz.*  $M(V_{b..n})$ . If there is *no* within group regression, we may then write

$$E_s(s_8^2) = \sigma_e^2; s_8^2 = \frac{S_8}{p-h}; S_8 = S_{bb..n} = p \cdot M(V_{b..n}) \text{ in Table 3} \quad \text{(xx)}$$

Thus we have

$$E_s(s_9^2) = \sigma_e^2; s_9^2 = S_8 - S_4 = r_{ab..m}^2 p \cdot M(V_{b..n}) \quad \text{(xxi)}$$

We have now all the requisite statistics of the battery of tests outlined above, and may proceed to define an  $F$ -ratio based on two statistics whose expected values are identical, if the null hypothesis is correct, choosing as the numerator the one whose expected value must be greater if the same hypothesis is false. The proof that they are independent Chi-Square variates follows the familiar lines set forth in Chapter 16 and in 17.04.

(i) Is there a group effect of either sort?

$$F_{35} = \frac{s_3^2}{s_5^2} = \frac{S_3}{p-2h} \frac{2h-2}{S_1-S_3} \quad \text{(xxii)}$$

(ii) Are the within-group regression coefficients identical?

$$F_{63} = \frac{s_6^2}{s_3^2} = \frac{S_4 - S_3}{h-1} \cdot \frac{p-2h}{S_3} \quad \text{(xxiii)}$$

(iii) If regression is uniform, is there an additive effect?

$$F_{47} = \frac{s_4^2}{s_7^2} = \frac{S_4}{p-h-1} \cdot \frac{h-1}{S_1-S_4} \quad \text{(xxiv)}$$

*Alternatively (for confirmation)*

$$F_{42} = \frac{s_4^2}{s_2^2} = \frac{S_4}{p-h-1} \frac{h-2}{S_2} \quad \text{(xxv)}$$



(iv) Is there regression not attributable to group effect?

$$F_{g_4} = \frac{s_9^2}{s_4^2} = \frac{(S_8 - S_4)(p - h - 1)}{S_4} \quad . \quad . \quad . \quad . \quad (\text{xxvi})$$

( $S_8 = S_{bb \cdot w}$  in Table 3).

*Note on Computations.* Table 3 is a key to the appropriate sums of square deviations and products of deviations from the relevant means. As in 17.02 above, it simplifies machine calculations if we sum 5 columns for each block of  $p_w$  paired scores :

$s_{a \cdot w}$	and	$s_{b \cdot w}$	total <i>raw scores</i> within block ;
$s_{aa \cdot w}$	and	$s_{bb \cdot w}$	total <i>squares</i> within block ;
$s_{ab \cdot w}$			total <i>products</i> within block.

We may write the corresponding grand totals for the  $p$  paired scores of *all* the  $h$  blocks as  $s_{a \cdot o}$ ,  $s_{b \cdot o}$ ,  $s_{aa \cdot o}$ ,  $s_{bb \cdot o}$  and  $s_{ab \cdot o}$ . We then have

$$S_{aa \cdot w} = s_{aa \cdot w} - \frac{s_{a \cdot w}^2}{p_w}; \quad S_{bb \cdot w} = s_{bb \cdot w} - \frac{s_{b \cdot w}^2}{p_w} \quad . \quad . \quad . \quad (\text{xxvii})$$

$$S_{aa..o} = s_{aa..o} - \frac{s_{a..o}^2}{p}; \quad S_{bb..o} = s_{bb..o} - \frac{s_{b..o}^2}{p} \quad . \quad . \quad . \quad (\text{xxviii})$$

$$S_{ab \cdot w} = s_{ab \cdot w} - \frac{s_{a \cdot w} \cdot s_{b \cdot w}}{p_w}; \quad S_{ab \cdot o} = s_{ab \cdot o} - \frac{s_{a \cdot o} \cdot s_{b \cdot o}}{p} \quad . \quad . \quad . \quad (xxix)$$

The statistics embodied in  $S_5$  are obtainable from the grid tautologies of 11.05 and 11.06, *viz.*:

$$V = M(V) + V(M); \text{Cov}(ab) = \text{Cov}(M_a M_b) + M \text{Cov}(ab).$$

Thus we have

$$S_{aa..m} = S_{aa..o} - S_{aa..p}; \quad S_{bb..m} = S_{bb..o} - S_{bb..p}. \quad . \quad . \quad (\text{xxx})$$

$$S_{ab \cdot m} = S_{ab \cdot o} - S_{ab \cdot p} \quad . \quad . \quad . \quad . \quad . \quad . \quad (\text{xxxi})$$

We may, however, use this relation as a check-up, if we compute directly

$$S_{aa \cdot m} = \sum_{w=1}^{w=h} \frac{s_{a \cdot w}^2}{p_w} - \frac{s_{a \cdot o}^2}{p} \quad . \quad . \quad . \quad . \quad . \quad . \quad (\text{xxxii})$$

$$S_{bb \cdot m} = \sum_{w=1}^{w=h} \frac{s_{b \cdot w}^2}{p_w} - \frac{s_{b \cdot 0}^2}{p} \quad . \quad . \quad . \quad . \quad . \quad . \quad (\text{xxxiii})$$

$$S_{a.b.m} = \sum_{w=1}^{w=h} \frac{S_{a.w} \cdot S_{b.w}}{p_w} - \frac{S_{a.o} \cdot S_{b.o}}{p} \quad . \quad . \quad . \quad . \quad (xxxiv)$$

*Numerical Example.* The table below exhibits three blocks each of six correlated variates  $a$  and  $b$  with corresponding squares and products for purposes of computation with totals at the foot.

TABLE 3

Group I					Group II					Group III				
$a$	$b$	$ab$	$a^2$	$b^2$	$a$	$b$	$ab$	$a^2$	$b^2$	$a$	$b$	$ab$	$a^2$	$b^2$
2	13	26	4	169	3	20	60	9	400	7	3	21	49	9
3	17	51	9	289	6	25	150	36	625	8	4	32	64	16
5	21	105	25	441	9	29	261	81	841	11	12	132	121	144
6	22	132	36	484	11	35	385	121	1225	13	17	221	169	289
8	25	200	64	625	12	37	444	144	1369	15	19	285	225	361
12	34	408	144	1156	13	37	481	169	1369	18	26	468	324	676
36	132	922	282	3164	54	183	1781	560	5829	72	81	1159	951	1495

Below we derive the appropriate sums as set out in Table 3 (Key for Computation) above.

Group I	Group II	Group III
$S_{aa.w} \quad . \quad . \quad 66$	$74$	$88$
$S_{bb.w} \quad . \quad . \quad 260$	$247.5$	$401.5$
$S_{ba.w} \quad . \quad . \quad 130$	$134$	$137$
$S_{aa.m} = 108$	$S_{bb.m} = 867$	$S_{ba.m} = -153$
$S_{aa.o} = 336$	$S_{bb.o} = 1776$	$S_{ba.o} = 298$
$S_{aa.p} = 228$	$S_{bb.p} = 909$	$S_{ba.p} = 451$

$$S_1 = 1776 - \frac{(298)^2}{336} \simeq 1512; \quad S_2 = 867 - \frac{(153)^2}{108} \simeq 650;$$

$$S_3 = \left[ 260 - \frac{(130)^2}{66} \right] + \left[ 247.5 - \frac{(134)^2}{74} \right] + \left[ 401.5 - \frac{(137)^2}{88} \right] \simeq 13;$$

$$S_4 = 909 - \frac{(451)^2}{228} \simeq 17; \quad p = 18; \quad h = 3.$$

Whence we have

$$(i) \quad F_{53} = \frac{1512 - 13 \cdot 12}{4} \cdot \frac{12}{13} \simeq 345.$$

$$(ii) \quad F_{63} = \frac{17 - 13 \cdot 12}{2} \cdot \frac{12}{13} \simeq 1.84.$$

$$(iii) \quad F_{74} = \frac{1512 - 17 \cdot 14}{2} \cdot \frac{14}{17} \simeq 616.$$

$$(iv) \quad F_{24} = \frac{650}{17} \cdot \frac{14}{1} \simeq 535.$$

The reader may find it instructive to investigate the approximate relationship which subsists between the  $a$  and  $b$  scores in each group (reference to the column totals provides a clue).



## 18.03 CAVEAT TO ANALYSIS OF COVARIANCE

In another context, we have had occasion to remind ourselves that statistical theory provides no sufficient substitute either for common sense or for an intimate knowledge of external nature variously denominated natural history, clinical experience, intuition and good judgment. It is especially important to keep this truth in full view, if we are to assess the value of analysis of covariance as a tool of research. In appropriate circumstances, the results of its application may be highly suggestive and helpful. It is not an *open sesame* to all closed doors between ignorance and knowledge when the end in view is to assess the relevance of quantitative and qualitative putative sources to a particular type of variation.

At the outset, it is necessary to emphasise (as stated in 18.01) that appropriate design of laboratory experiments (as opposed to field trials and industrial experimentation so-called) commonly offers a more direct and satisfactory approach to the issue which is the peculiar concern of the technique under discussion. While there may admittedly exist circumstances which make a putatively relevant quantitative source of contributory variation difficult or even impossible to control in an experiment conducted to evaluate the significance of a second and qualitative criterion of classification, it is also true that such circumstances commonly exclude the possibility of taking precautions to assess the validity of the assumptions inherent in the method of 18.02.

The tests dealt with in 18.02 are *conditional* on two assumptions. One is that regression is linear. The other is that the residual variation of which  $\sigma_e^2$  is the measure is the same for all sub-universes. The second we can test, if in doubt, by methods mentioned elsewhere. Indeed, the test for uniformity of regression answers the question, as far as it is possible to give an answer to it, if there is no reason to dismiss the hypothesis that there is uniformity of regression. The assumption of linearity is *not* one which we can commonly and conclusively justify in situations which compel us to fall back on the analysis of covariance as an alternative to a more direct procedure. The reason for this is one we have noted elsewhere (pp. 741 and 742) *en passant*. The linearity test of 17.04 breaks down, unless we can arrange matters so that we have more than one *B*-score value for at least some of the *A*-scores, as we can ensure in certain types of experimental design. In the type dealt with in 18.02, we have in fact to take our *B*-scores as they come.

For both the reasons last stated, it is important to be quite clear about the credentials of the claim that statistical tests such as the Gosset *t*-test and any test based on an *F*-ratio permit us to make assertions with confidence about small samples. Formally, and in accordance with our initial assumptions (e.g. normally distributed scores), it is true to say that such tests, unlike tests in common use a generation since, rely on the distribution of sample values of the parameters of the relevant distribution, in contradistinction to the distribution of ratios involving the unknown parameters which we can at best estimate with assurance for very large samples. On the other hand, it is necessary to remind ourselves that a significance test of the sort under discussion can merely give us a rule for dismissing a null hypothesis without risk of doing so wrongly very often. It cannot give us good reasons for believing it, though we may indeed have derived reasons from other sources; and a sample, if small enough, may give the test rule little chance of dismissing a null hypothesis which is false.

To the author, the moral of this is clear. If we have before us large sub-samples in the set-up of 18.02, we have good enough reason to justify the conviction that regression of the *B*-score on the *A*-score is approximately linear. In any case, our data may be such as to exclude the possibility of checking this assumption by recourse to an appropriate significance test; and in any case, the test does not conclusively prove that the null hypothesis is correct. In



practice, therefore, analysis of covariance or any other technique conditional on an assumption (e.g. linearity) which we have not good reason to adopt for reasons other than the outcome of a statistical test, cannot base its legitimate claim to consideration on the economy of working with small samples.

For a reason stated in 17.06, it may also be legitimate to express some doubt about the wisdom of assuming that the analysis of covariance is a reliable tool of research in the *concurrent* domain of sociology and econometrics. Within the *consequential* domain of agricultural field trials or nutritional science, the meaning of the significance tests is clear, and the implications of the principle of the fixed-*A* set—Churchill Eisenhart's Model I approach (p. 548)—present no semantic difficulties. It is not equally clear what the Model I approach signifies in the unique historical situations of sociological enquiry.

One other consideration bearing on the judicious use of Analysis of Covariance calls for comment; and is on all fours with a limitation perhaps too little emphasised in connexion with the parent technique of Chapter 13. We may speak of it as the *dilution of the class effect*. The method of 18.02 is, of course, applicable to situations in which we distinguish only two classes of paired scores. When the number of classes is large, there is always a possibility that the effect of others will conceal one which is out of step. For this reason, it is a wise precaution to calculate within-group correlation coefficients for comparison, and separate investigation, if the figures are suggestive.

#### 18.04 THE CONCEPT OF FACTOR PATTERN

In different contexts we use the word statistics in several ways, the connexions between one and the other use being various and somewhat exiguous. Originally, it signified numerical information bearing on the affairs of *State*; and it is well to remind ourselves of this when we examine the controversies provoked by the applications of *factor analysis* in the field of psychology. Factor analysis is a statistical procedure largely developed in connection with what is basically an administrative issue, i.e. *personnel selection*. To trace its origins we must go back to work undertaken in the nineties by Binet and Simon with a view to devising tests of intellectual aptitude with more prognostic value than scholastic examinations. At the start, the yardstick of such so-called intelligence tests was their correspondence with teachers' estimates of relative ability. As we have seen in Chapter 6 of Vol. I, this is essentially a problem of rank correlation. The pioneers of test-work explored from this standpoint a variety of puzzles, the general knowledge quiz and feats of memory presumptively unrelated to school training. There thus emerged a mass of information about the intercorrelation of test scores.

Intercorrelation of two test scores in this setting signifies a correlation—rank or product-moment—based on the application of each of them to each individual of a group. In the idiom of our 2-face card pack model, of 12.00, the individual is the card and the test score is the number of pips on one or other face. In that of the Umpire Bonus Model, the individual is the particular *trial* and the test scores are the scores of the players. The fact that the results of say six tests of the same individual tie-up in the sense that there is a significant positive correlation for any pair of corresponding scores applied to the same, or to a comparable, group of individuals does **not** necessarily mean that they all measure the same aptitude. This will still be true if: (a) each test measures a congeries of unconnected attributes; (b) the results of any pair of them depend in part on one such attribute which does not affect those of any remaining test. Hence arises the following question: is there any characteristic common to what we are measuring when we apply different so-called intelligence tests? If so, and if there is indeed only one such



sort of aptitude, it may be a verbal convenience to adopt the convention of restricting the use of the word intelligence thereto.

The query last stated assumed a more provocative aspect when Spearman first propounded what it is customary to call the *hierarchical* principle embodying a feature of the lay-out we may meet when we arrange such test-score intercorrelations in a symmetrical grid. His interpretation embodied in the concept denoted by  $g$  (for *general* intelligence) started a controversy which led to the recognition of more complicated patterns. Factor analysis is the attempt to interpret them. At the outset then, let us be clear about what a *factor pattern* signifies. We shall assume that we have before us the results of applying 5 tests ( $A-E$ ) to each boy or girl in a school form, and that we can therefore calculate the product-moment index of any particular pair. Having done so we may set them out gridwise as below on the left. To illustrate the meaning of the simplest kind of hierarchical pattern, we shall suppose that the numerical values are as shown on the right.

	$A$	$B$	$C$	$D$	$E$
$A$	—	$r_{ab}$	$r_{ac}$	$r_{ad}$	$r_{ae}$
$B$	$r_{ab}$	—	$r_{bc}$	$r_{bd}$	$r_{be}$
$C$	$r_{ac}$	$r_{bc}$	—	$r_{cd}$	$r_{ce}$
$D$	$r_{ad}$	$r_{bd}$	$r_{cd}$	—	$r_{de}$
$E$	$r_{ae}$	$r_{be}$	$r_{ce}$	$r_{de}$	—

	$A$	$B$	$C$	$D$	$E$
$A$	—	0.44	0.61	0.53	0.73
$B$	0.44	—	0.36	0.29	0.39
$C$	0.61	0.36	—	0.43	0.55
$D$	0.53	0.29	0.43	—	0.50
$E$	0.73	0.39	0.55	0.50	—

A close inspection of the figures shows that there is a rank correspondence, which comes into focus when we rearrange the items as below on the left. This hierarchical correspondence is not the only circumstance the figures suggest. Closer inspection shows that the ratio of items in any two rows of one column is roughly the same as that of corresponding items of the same two rows in another column. That this is so becomes evident, if we assign suitable border factors to each corresponding row and column, expressing the cell entries as their products. Thus we can closely reproduce the actual *correlation matrix* on the left below by multiplying factors assigned to heads of columns and row margins on the right:

	$A$	$E$	$C$	$D$	$B$		0.9	0.8	0.7	0.6	0.5
$A$	—	0.73	0.61	0.53	0.44	0.9	—	0.72	0.63	0.54	0.45
$E$	0.73	—	0.55	0.50	0.39	0.8	0.72	—	0.56	0.48	0.40
$C$	0.61	0.55	—	0.43	0.36	0.7	0.63	0.56	—	0.42	0.35
$D$	0.53	0.50	0.43	—	0.29	0.6	0.54	0.48	0.42	—	0.30
$B$	0.44	0.39	0.36	0.29	—	0.5	0.45	0.40	0.35	0.30	—

The lay-out of the figures of this fictitious example illustrates the simplest type of *factor pattern*, i.e. the *hierarchical* or single factor pattern for which Spearman first offered a theoretical interpretation. How later test results suggest more complex patterns the following actual results cited by Burt will serve to illustrate:

	<i>Composition</i>	<i>Handicraft</i>	<i>Spelling</i>	<i>Drawing</i>	<i>Reading</i>	<i>Writing</i>
<i>Composition</i>	—	0.30	0.49	0.38	0.58	0.44
<i>Handicraft</i>	0.30	—	0.09	0.50	0.10	0.28
<i>Spelling</i>	0.49	0.09	—	0.12	0.46	0.25
<i>Drawing</i>	0.38	0.50	0.12	—	0.13	0.36
<i>Reading</i>	0.58	0.10	0.46	0.13	—	0.21
<i>Writing</i>	0.44	0.28	0.25	0.36	0.21	—

A pattern, suggestive of a hierarchical relationship between a sub-group of this battery of tests, emerges when we rearrange the figures as below :

	<i>Composition</i>	<i>Reading</i>	<i>Spelling</i>	<i>Writing</i>	<i>Drawing</i>	<i>Handicraft</i>
<i>Composition</i>	—	0.58	0.49	0.44	0.38	0.30
<i>Reading</i>	0.58	—	0.46	0.21	0.13	0.10
<i>Spelling</i>	0.49	0.46	—	0.25	0.12	0.09
<i>Writing</i>	0.44	0.21	0.25	—	0.36	0.28
<i>Drawing</i>	0.38	0.13	0.12	0.36	—	0.50
<i>Handicraft</i>	0.30	0.10	0.09	0.28	0.50	—

To what extent it is possible to interpret such a pattern as the above with confidence, we shall discuss at a later stage. First, let us examine the rationale of Spearman's hierarchical criterion of the single factor pattern. Spearman himself, and many who have followed him, relied on reasoning which involves an unnecessary and not necessarily true limitation, *viz.* the assumption that linear regression is itself a necessary consequence of linear concomitant variation. On the other hand, our exploration of the Umpire Bonus Model in 12.01 opens the door to a very simple derivation of the hierarchical criterion, if we assume with Spearman that the test score of an individual is the algebraic sum of independent hypothetical components. The single factor pattern for test scores  $x_a$ ,  $x_b$ , etc., then takes the following form,  $x_u$  being the common component and  $x_{a.o}$ ,  $x_{b.o}$ , etc., the specific one :

$$x_a = A_u \cdot x_u + A_o \cdot x_{a.o};$$

$$x_b = B_u \cdot x_u + B_o \cdot x_{b.o};$$

$$x_c = C_u \cdot x_u + C_o \cdot x_{c.o}.$$

Strictly speaking, the assumption of the statistical independence of the two score components is unnecessary. All we require to assume is that their covariance is zero. The same is true of the score components in the Umpire Bonus Model set-up.

#### 18.05 DERIVATION OF THE HIERARCHICAL CRITERION

The foregoing system of equations definitive of the individual test score are formally identical with (vi) in 12.01, the common factor being the umpire's bonus. For this set-up, we have established the following relation specified by (ix) of 12.01 :

$$r_{ab} = r_{au} r_{bu}.$$



For a set of 5 players we can therefore write

	$r_{au}$	$r_{bu}$	$r_{cu}$	$r_{du}$	$r_{eu}$
$r_{au}$	...	$r_{ab} = r_{au} \cdot r_{bu}$	$r_{ac} = r_{au} \cdot r_{cu}$	$r_{ad} = r_{au} \cdot r_{du}$	$r_{ae} = r_{au} \cdot r_{eu}$
$r_{bu}$	$r_{ab} = r_{au} \cdot r_{bu}$	...	$r_{bc} = r_{bu} \cdot r_{cu}$	$r_{bd} = r_{bu} \cdot r_{du}$	$r_{be} = r_{bu} \cdot r_{eu}$
$r_{cu}$	$r_{ac} = r_{au} \cdot r_{cu}$	$r_{bc} = r_{bu} \cdot r_{cu}$	...	$r_{cd} = r_{cu} \cdot r_{du}$	$r_{ce} = r_{cu} \cdot r_{eu}$
$r_{du}$	$r_{ad} = r_{au} \cdot r_{du}$	$r_{bd} = r_{bu} \cdot r_{du}$	$r_{cd} = r_{cu} \cdot r_{du}$	...	$r_{de} = r_{du} \cdot r_{eu}$
$r_{eu}$	$r_{ae} = r_{au} \cdot r_{eu}$	$r_{be} = r_{bu} \cdot r_{eu}$	$r_{ce} = r_{cu} \cdot r_{eu}$	$r_{de} = r_{du} \cdot r_{eu}$	...

Given that all five players receive a *single* bonus from one and the same umpire, we may thus lay out the correlation matrix of the intercorrelations between their scores in such a way that each cell entry is the product of the two border *factors* respectively definitive of the  $p$ - $m$  coefficient of the umpire's own score with that of one or other player. This establishes the conclusion that the single factor hypothesis is a *sufficient* condition of the hierarchical pattern; but we have still to show that it is a *necessary* one, i.e. the only feasible explanation.

Before exploring this issue, let us notice a principle inherent in the hierarchical criterion. If we consider any four players of a set of this sort, we notice that we can pair them off thus  $AB \cdot CD$ ;  $AC \cdot BD$ ;  $AD \cdot BC$ . The products of the intercorrelation of the scores of one pair with that of the residual pair are then as follows:

$$r_{ab} \cdot r_{cd} = A_u B_u \frac{\sigma_u^2}{\sigma_a \sigma_b} \cdot C_u D_u \frac{\sigma_u^2}{\sigma_c \sigma_d} = \frac{A_u B_u C_u D_u V_u^2}{\sigma_a \sigma_b \sigma_c \sigma_d};$$

$$r_{ac} \cdot r_{bd} = A_u C_u \frac{\sigma_u^2}{\sigma_a \sigma_c} \cdot B_u D_u \frac{\sigma_u^2}{\sigma_b \sigma_d} = \frac{A_u B_u C_u D_u V_u^2}{\sigma_a \sigma_b \sigma_c \sigma_d};$$

$$r_{ad} \cdot r_{bc} = A_u D_u \frac{\sigma_u^2}{\sigma_a \sigma_d} \cdot B_u C_u \frac{\sigma_u^2}{\sigma_b \sigma_c} = \frac{A_u B_u C_u D_u V_u^2}{\sigma_a \sigma_b \sigma_c \sigma_d}.$$

Hence we arrive at Spearman's *tetrad equations*, viz.:

$$r_{ab} \cdot r_{cd} - r_{ac} \cdot r_{bd} = 0;$$

$$r_{ab} \cdot r_{cd} - r_{ad} \cdot r_{bc} = 0;$$

$$r_{ac} \cdot r_{bd} - r_{ad} \cdot r_{bc} = 0.$$

Thus a single factor pattern implies that for any 4-fold set of test scores the *tetrad differences do not significantly differ from zero*. In deriving this result we have assumed a strictly linear law of the composition of the scores. Let us now examine the consequences of a non-linear law for the same model, viz.:

$$X_a = A_u X_u^p + A_o X_{a.o};$$

$$X_b = B_u X_u^q + B_o X_{b.o};$$

$$X_c = C_u X_u^s + C_o X_{c.o};$$

$$X_d = D_u X_u^t + D_o X_{d.o}.$$

We may then write

$$\begin{aligned} \text{Cov}(x_a, x_b) &= E(X_a^p \cdot X_b^q) = A_u B_u E(X_u^{p+q}) \\ &= A_u B_o E(X_u^p \cdot X_{b.o}) + A_o B_u E(X_{a.o} \cdot X_u^q) + A_o B_o E(X_{a.o} \cdot X_{b.o}). \end{aligned}$$

If there is zero covariance between all powers of the two score components and  $m_{p+q}$  is the  $(p+q)$ th mean moment of the umpire's score distribution, we therefore have

$$\text{Cov}(x_a, x_b) = A_u B_u E(X_u^{p+q}) = A_u B_u \cdot m_{p+q}.$$

Similarly,

$$\text{Cov}(x_a, x_c) = A_u C_u \cdot m_{p+s}, \text{ etc.}$$

We thus derive

$$r_{ab} \cdot r_{cd} = \frac{A_u B_u C_u D_u}{\sigma_a \sigma_b \sigma_c \sigma_d} \cdot m_{p+q} \cdot m_{s+t};$$

$$r_{ac} \cdot r_{bd} = \frac{A_u B_u C_u D_u}{\sigma_a \sigma_b \sigma_c \sigma_d} \cdot m_{p+s} \cdot m_{q+t};$$

$$r_{ad} \cdot r_{bc} = \frac{A_u B_u C_u D_u}{\sigma_a \sigma_b \sigma_c \sigma_d} \cdot m_{p+t} \cdot m_{q+s}.$$

Thus the tetrad differences will vanish if, and only if,

$$m_{p+q} \cdot m_{s+t} = m_{p+s} \cdot m_{q+t} = m_{p+t} \cdot m_{q+s}.$$

This will be true if  $p = q = s = t$ , in which case the law of composition is linear since we may replace  $X_u^p, X_u^q$ , etc., by  $Z_u$ .

If two umpires ( $U$  and  $W$ ) contribute to the total score of each player, (xii) of 12.01 prescribes that

$$r_{ab} = r_{au} \cdot r_{bu} + r_{aw} \cdot r_{bw}, \text{ etc.};$$

$$r_{au} = A_u \frac{\sigma_u}{\sigma_a}; \quad r_{aw} = A_w \frac{\sigma_w}{\sigma_a}, \text{ etc.}$$

Whence we derive

$$r_{ab} \cdot r_{cd} = (r_{au} \cdot r_{bu} + r_{aw} \cdot r_{bw})(r_{cu} \cdot r_{du} + r_{cw} \cdot r_{dw}), \text{ etc.}$$

Our tetrads thus reduce to

$$r_{ab} \cdot r_{cd} = r_{au} r_{bu} r_{cu} r_{du} + r_{aw} r_{bw} r_{cw} r_{dw} + r_{au} r_{bu} r_{cw} r_{dw} + r_{aw} r_{bw} r_{cu} r_{du};$$

$$r_{ac} \cdot r_{bd} = r_{au} r_{bu} r_{cu} r_{du} + r_{aw} r_{bw} r_{cw} r_{dw} + r_{au} r_{bw} r_{cu} r_{dw} + r_{aw} r_{bu} r_{cw} r_{du};$$

$$r_{ad} \cdot r_{bc} = r_{au} r_{bu} r_{cu} r_{du} + r_{aw} r_{bw} r_{cw} r_{dw} + r_{au} r_{bw} r_{cw} r_{du} + r_{aw} r_{bu} r_{cu} r_{dw}.$$

The first two terms in each of the above are identical. So the tetrad differences will vanish only if

$$\begin{aligned} r_{au} \cdot r_{bu} \cdot r_{cw} \cdot r_{dw} + r_{aw} \cdot r_{bw} \cdot r_{cu} \cdot r_{du} &= r_{au} \cdot r_{cu} \cdot r_{bw} \cdot r_{dw} + r_{aw} \cdot r_{cw} \cdot r_{bu} \cdot r_{du} \\ &= r_{au} \cdot r_{du} \cdot r_{bw} \cdot r_{cw} + r_{aw} \cdot r_{dw} \cdot r_{bu} \cdot r_{cu}. \end{aligned}$$

From the first pair we get

$$r_{au} \cdot r_{dw}(r_{bu} \cdot r_{cw} - r_{bw} \cdot r_{cu}) = r_{aw} \cdot r_{du}(r_{bu} \cdot r_{cw} - r_{bw} \cdot r_{cu}),$$

$$\therefore \frac{r_{au}}{r_{aw}} = \frac{r_{du}}{r_{dw}}.$$

By pairing off each of the three identities, we thus get

$$\frac{r_{au}}{r_{aw}} = \frac{r_{bu}}{r_{bw}} = \frac{r_{cu}}{r_{cw}} = \frac{r_{du}}{r_{dw}} = K.$$



Hence we may write

$$r_{ab} = r_{aw} \cdot r_{bw}(1 + K^2), \text{ etc.}$$

This will be true if

$$\begin{aligned} X_a &= A_w(1 + K^2)^{\frac{1}{2}}X_w + A_o \cdot X_{a.o}; & X_b &= B_w(1 + K^2)^{\frac{1}{2}}X_w + B_o \cdot X_{b.o}, \text{ etc.} \\ A_u X_u + A_w X_w &= A_u(1 + K^2)^{\frac{1}{2}}X_w; & B_u X_u + B_w X_w &= B_w(1 + K^2)^{\frac{1}{2}}X_w, \text{ etc.} \end{aligned}$$

From an algebraic viewpoint, we may therefore say that the tetrad identities are valid for a set-up involving 2 umpires only if the composite bonus is replaceable by that of either alone with appropriate change of scale. This resolves a controversy concerning how far we are entitled to regard Spearman's  $g$  as a single entity. The answer is that it behaves as such in the Spearman framework of test material, in the sense that the carbon atom behaves as a unit in the customary manipulations of chemical analysis. This does not exclude the possibility that it might behave otherwise in a different framework as does the carbon atom under the impact of radioactive emanations.

With the qualification last stated, we can say that the single factor postulate is both a sufficient and necessary condition of a strictly hierarchical pattern, as defined in 18.04. If we now fill in the empty diagonal cells of our correlation matrix by the appropriate entries, *viz.*:  $r_{au}^2$ ,  $r_{bu}^2$ , etc., each such statistic has an intelligible meaning *vis-à-vis* our bonus model, i.e. the correlation of the player's score with that of the umpire. We have yet to give these diagonal cell entries a meaning in the domain of factor analysis; and we can do so, if we go back to our model, recalling that

$$\begin{aligned} V_a &= A_u^2 \cdot V_u + A_o^2 \cdot V_{a.o}; \\ r_{au}^2 &= \frac{A_u^2 \cdot V_u}{V_a}. \end{aligned}$$

The first of these two equations exhibits the breakdown of the variance of the  $A$ -score distribution into 2 moieties, one we may call  $V_o$ , the part referable to the bonus which the players have in common, and one  $V_s$  the part referable to what each player records as his individual score. Thus the proportionate contribution of the bonus to the total variance is

$$\frac{V_o}{V_a} = \frac{A_u^2 \cdot V_u}{V_a} = r_{au}^2.$$

In terms of factor analysis, each hypothetical diagonal entry of the test-score correlation matrix when completed is therefore the proportionate contribution of the hypothetical common factor to the variance of a test-score distribution. To proceed further, we may simplify our problem by assuming that our test scores are standard scores as in 12.07, and that the total variance of the test-score distribution is therefore unity.

#### 18.06 RELIABILITY, ATTENUATION AND COMMUNALITY

In the real world, no performance test is perfect in the sense that successive applications to the same individual would yield absolutely identical score values; and we may take stock of the implications of the imperfect reliability of any test by a simple physical analogy to the Umpire Bonus set-up. Within certain limits the law relating the load to the extension of a spring is very closely linear, and we may suppose that we apply to each of two springs  $A$  and  $B$  the same load at the same trial and different loads at different trials, the test score ( $x_a$ ,  $x_b$ ) being the length of the spring. To apply the test we use a vernier which is fallible like all instruments, and we

may therefore regard the test score as having 2 components: (i) its true value proportional to the applied force ( $x_f$ ) which varies from trial to trial in accordance with Hooke's law; (ii) a random component  $x_e$ ; so that

$$x_a = A_f \cdot x_f + x_{ea} \text{ and } x_b = B_f \cdot x_f + x_{eb}.$$

If we use the same vernier, the variance of the distribution of the error component is the same throughout, and

$$r_{ab}^2 = \frac{A_f^2 B_f^2 V_f}{(A_f^2 V_f + V_e)(B_f^2 V_f + V_e)}.$$

With this analogy to the Umpire Bonus set-up in mind, we may now formulate the components of two composite scores recorded for the same group of persons in the following terms:

- (a) one ( $x_f$ ) which assesses a common attribute;
- (b) one ( $x_{sa}$  or  $x_{sb}$ ) which assesses a specific attribute;
- (c) one ( $x_{ea}$ ,  $x_{eb}$ ) which refers to error of observation, or what is indistinguishable therefrom, i.e. uncontrolled circumstances affecting the response of the test subject.

We then write our scoring system in terms of score deviations as

$$X_a = A_f \cdot X_f + A_s \cdot X_{sa} + X_{ea};$$

$$X_b = B_f \cdot X_f + B_s \cdot X_{sb} + X_{eb}.$$

The error and specific components have by hypothesis zero covariance, and they will be indistinguishable if we apply once only each composite test to each member of the test group of persons, so that we might then write

$$A_o \cdot x_a \cdot o = A_s \cdot x_{sa} + x_{ea} \text{ and } B_o \cdot x_b \cdot o = B_s \cdot x_{sb} + x_{eb}$$

for the components peculiar to each test. Otherwise, we must write

$$\therefore r_{ab}^2 = \frac{A_f^2 B_f^2 V_f^2}{(A_f^2 V_f + A_s^2 V_{sa} + V_{ea})(B_f^2 V_f + B_s^2 V_{sb} + V_{eb})} \quad \cdot \quad \cdot \quad \cdot \quad (i)$$

In the absence of error, we might write the true correlation ( $r_{ab \cdot e}$ ) in accordance with the implicit assumption  $V_{ea} = 0 = V_{eb}$ , so that

$$\begin{aligned} r_{ab \cdot e}^2 &= \frac{A_f^2 B_f^2 V_f^2}{(A_f^2 V_f + A_s^2 V_{sa})(B_f^2 V_f + B_s^2 V_{sb})}, \\ \frac{r_{ab \cdot e}^2}{r_{ab}^2} &= \frac{A_f^2 V_f + A_s^2 V_{sa} + V_{ea}}{A_f^2 V_f + A_s^2 V_{sa}} \cdot \frac{B_f^2 V_f + B_s^2 V_{sb} + V_{eb}}{B_f^2 V_f + B_s^2 V_{sb}} \quad \cdot \quad \cdot \quad \cdot \quad (ii) \end{aligned}$$

Let us now suppose that we repeat the same test on each person, so that we may correlate the first composite test score of each person with his or her second. In this set-up  $x_{sa}$  and  $x_f$  are each common components of the two scores at the same trial (i.e. the two scores of the same person); but the error component differs from trial to trial. For test  $A$  we may therefore write our equations as

$$x_{a1} = A_f x_f + A_{sa} x_{sa} + x_{e1};$$

$$x_{a2} = A_f x_f + A_{sa} x_{sa} + x_{e2}.$$

The situation so described is one in which  $x_f$  and  $x_{sa}$  correspond to the independent contributions of different umpires and  $x_{e1}$ ,  $x_{e2}$  correspond to the individual components  $x_{a \cdot o}$  and  $x_{b \cdot o}$  of our bonus model. On the assumption that errors of a given magnitude occur with



equal frequency on both occasions  $V_{e1} = V_{ea} = V_{e2}$ . Thus we may write the product-moment coefficient of the *test-retest* as

$$r_{aa} = \frac{A_f^2 V_f + A_s^2 V_{sa}}{A_f^2 V_f + A_s^2 V_{sa} + V_{ea}} = \frac{A_f^2 V_f + A_s^2 V_{sa}}{V_a} \quad . \quad . \quad . \quad (iii)$$

Similarly, we may write

$$r_{bb} = \frac{B_f^2 V_f + B_s^2 V_{sb}}{B_f^2 V_f + B_s^2 V_{sb} + V_{eb}} = \frac{B_f^2 V_f + B_s^2 V_{sb}}{V_b} \quad . \quad . \quad . \quad (iv)$$

By substitution of (iii) and (iv) in (ii) we therefore obtain the so-called *correction formula for attenuation*, viz. :

$$r_{ab.e}^2 = \frac{r_{ab}^2}{r_{aa} \cdot r_{bb}} \quad \text{or} \quad r_{ab.e} = \frac{r_{ab}}{\sqrt{r_{aa} \cdot r_{bb}}} \quad . \quad . \quad . \quad (v)$$

In this formula,  $r_{ab.e}$  is the expected value of the product-moment correlation between  $A$  and  $B$  test scores in the absence of errors incident to the carrying out of the test,  $r_{ab}$  being its crude value. We may speak of  $r_{aa}$  and  $r_{bb}$  as *coefficients of reliability*, since their value is unity if there is perfect test-retest agreement, in which case  $x_e$  is constant and  $V_e$  is zero.

In the set-up under consideration, we may label the components of our test-score deviations, thus

TEST SCORE	Communality	Specificity	Error
$X_a =$	$A_f X_f$	$+ A_s X_{sa}$	$+ A_e X_{ea}$
$X_b =$	$B_f X_f$	$+ B_s X_{sb}$	$+ B_e X_{eb}$
	RELIABILITY		ERROR

In this set-up the correlation between tests  $A$  and  $B$  assessed by the  $p$ - $m$  index ( $r_{ab}$ ) of the test scores  $x_a, x_b$ , is comparable with the correlation of the scores of two players who receive multiples of the same bonus from one umpire in the model of 12.01. If we could actually isolate the common factor ( $x_f$ ) equivalent to the umpire's bonus, we could then record  $r_{af}$  or  $r_{bf}$  the test-score factor correlation corresponding to the correlation between the umpire's score and that of the player. Formally, this is

$$r_{af}^2 = \frac{A_f^2 V_f}{V_a} \quad \text{and} \quad r_{bf}^2 = \frac{B_f^2 V_f}{V_b}.$$

The corresponding components of variance for test  $A$  are thus :

Source	Actual	Proportionate
<i>Total</i>	$V_a$	1
<i>Communality</i>	$A_f^2 V_f$	$r_{af}^2$
<i>Specificity</i>	$A_s^2 V_{sa}$	$r_{aa} - r_{af}^2$
<i>Reliability</i>	$A_f^2 V_f + A_s^2 V_{sa}$	$r_{aa}$
<i>Error</i>	$A_e^2 V_{ae}$	$1 - r_{aa}$

Let us now return to our equations exhibiting the three additive components, and express them in standard form, i.e.

$$z_a = \frac{X_a}{\sigma_a}, \text{ etc.}$$

Thus we may put

$$\frac{X_a}{\sigma_a} = \frac{A_f \sigma_f}{\sigma_a} \cdot \frac{X_f}{\sigma_f} + \frac{A_s \sigma_{sa}}{\sigma_a} \cdot \frac{X_{sa}}{\sigma_{sa}} + \frac{A_e \sigma_{ea}}{\sigma_a} \cdot \frac{X_{ea}}{\sigma_{ea}}.$$

We may then write

$$\frac{A_f \sigma_f}{\sigma_a} = a_f; \quad \frac{A_s \sigma_{sa}}{\sigma_a} = a_s; \quad \frac{A_e \sigma_{ea}}{\sigma_a} = a_e.$$

The definitive equation of the  $A$  test score thus becomes

$$z_a = a_f z_f + a_s z_{sa} + a_e z_{ea}.$$

The variance of the distribution of our standard scores being unity, we may therefore write

$$a_f^2 + a_s^2 + a_e^2 = 1;$$

$$a_f^2 = \frac{A_f^2 \cdot V_f}{V_a} = r_{af}^2;$$

$$a_f^2 + a_s^2 = \frac{A_f^2 \cdot V_f}{V_a} + \frac{A_s^2 \cdot V_{sa}}{V_a} = r_{aa}.$$

Whence we may write the proportionate contributions to total variance as

<i>Communality</i>	$a_f^2 = r_{af}^2;$
<i>Reliability</i>	$a_f^2 + a_s^2 = r_{aa};$
<i>Specificity</i>	$a_s^2 = r_{aa} - r_{af}^2;$
<i>Error</i>	$a_e^2 = 1 - r_{aa}.$

Here a word of caution may not be amiss. It might seem more consistent with the pattern of the correlation matrix to denote by  $r_{aa}$ ,  $r_{bb}$ , etc. the entries in the left-right-downwards diagonal. We correctly label them in the above symbolism as  $r_{af}^2$ ,  $r_{bf}^2$ , etc. In the jargon of factor analysis we speak of  $a_f = r_{af}$  as a *factor loading* or *saturation*. When there are several common factors contributing to a set of intercorrelations, we may number them as below:

$$z_a = a_1 z_1 + a_2 z_2 + a_3 z_3 \dots a_s z_{sa} + a_e z_{ea};$$

$$z_b = b_1 z_1 + b_2 z_2 + b_3 z_3 \dots b_s z_{sb} + b_e z_{eb}.$$

It follows from results obtained in 12.01 that

$$r_{a1} = a_1; \quad r_{a2} = a_2, \text{ etc.}$$

The total communalities w.r.t. the two tests are then

$$a_1^2 + a_2^2 + a_3^2 \dots \text{ and } b_1^2 + b_2^2 + b_3^2 \dots$$

The  $p$ - $m$  intercorrelation for tests  $A$  and  $B$  is

$$a_1 b_1 + a_2 b_2 + a_3 b_3 \dots$$



## 18.07 THE SINGLE FACTOR PATTERN

We have seen that a truly hierarchical pattern embodies the set of identities called vanishing tetrad differences. Before we explore a way of determining the factor loadings for such a pattern, it will be useful to recognise another set of identities involving *triads*, e.g.  $r_{ab}$ ,  $r_{ac}$  and  $r_{bc}$ . For the single factor pattern in the symbols of 18.05,

$$r_{ab} \cdot r_{ac} = (r_{au} \cdot r_{bu}) (r_{au} \cdot r_{cu}) = r_{au}^2 r_{bu} r_{cu} = r_{au}^2 \cdot r_{bc},$$

$$\therefore \frac{r_{ab} \cdot r_{ac}}{r_{bc}} = r_{au}^2 \quad . \quad . \quad . \quad . \quad . \quad . \quad (i)$$

Similarly we derive

$$\frac{r_{ab} \cdot r_{ad}}{r_{bd}} = r_{au}^2; \quad \frac{r_{ac} \cdot r_{ad}}{r_{cd}} = r_{au}^2 \quad . \quad . \quad . \quad . \quad . \quad . \quad (ii)$$

From any four intercorrelations involving test  $A$  we can thus derive 3 estimates of the communality  $r_{au} = a_u$  and similarly 3 of  $b_u$ ,  $c_u$ ,  $d_u$ , etc. It follows that the triad ratios defined by (i) and (ii) cannot significantly exceed unity and should not significantly differ *inter se* if the product in the numerator has the same common factor for all. In practice, however, we shall not expect perfect agreement. For the example of a good hierarchical pattern in 18.04 (p. 785) the triads involving the  $A$  test communality are  $ABC$ ,  $ABD$ ,  $ABE$ ,  $ACD$ ,  $ACE$ ,  $ADE$ . These yield the following numerical values of  $r_{au}^2$ : 0.75, 0.80, 0.83, 0.75, 0.81, 0.77. This raises the question: how can we choose a satisfactory average? We may write the sum of the numerator products of the foregoing set of triads in the 5-test table as

$$S_{a(ij)} = r_{ab} \cdot r_{ac} + r_{ab} \cdot r_{ad} + r_{ab} \cdot r_{ae} + r_{ac} \cdot r_{ad} + r_{ac} \cdot r_{ae} + r_{ad} \cdot r_{ae}.$$

From the derivation of (i) and (ii) we see that

$$S_{a(ij)} = r_{au}^2 (r_{bu} \cdot r_{cu} + r_{bu} \cdot r_{du} + r_{bu} \cdot r_{eu} + r_{cu} \cdot r_{du} + r_{cu} \cdot r_{eu} + r_{du} \cdot r_{eu}).$$

Similarly, we may write the sum of terms corresponding to the denominators of (i) or (ii) as

$$\begin{aligned} S_{ij} &= r_{bc} + r_{bd} + r_{be} + r_{cd} + r_{ce} + r_{de} \\ &= r_{bu} \cdot r_{cu} + r_{bu} \cdot r_{du} + r_{bu} \cdot r_{eu} + r_{cu} \cdot r_{du} + r_{cu} \cdot r_{eu} + r_{du} \cdot r_{eu}. \end{aligned}$$

Whence we have

$$r_{au}^2 = \frac{S_{a(ij)}}{S_{ij}} \quad . \quad . \quad . \quad . \quad . \quad . \quad (iii)$$

In this expression  $S_{a(ij)}$  means: *add up all the products of two different cell entries in either the A-row or the A-column of the correlation matrix.* The denominator  $S_{ij}$  means: *add all the entries*

on one side of the left-right-downwards diagonal, excluding those in both the *A*-row and the *A*-column. For the specimen set-up on p. 785 (18.04) we have :

<i>A</i> <i>B</i> <i>C</i> <i>D</i> <i>E</i>					Product Total $S_{a(ii)}$
<i>A</i> ...	$r_{ab}$	$r_{ac}$	$r_{ad}$	$r_{ae}$	$r_{ab}r_{ac} + r_{ab}r_{ad} + r_{ab}r_{ae} + r_{ac}r_{ad}$ , etc.
<i>B</i> ...	...	...	...	...	
<i>C</i> ...	$r_{bc}$	...	...	...	
<i>D</i> ...	$r_{bd}$	$r_{cd}$	...	...	
<i>E</i> ...	$r_{be}$	$r_{ce}$	$r_{de}$	...	
Total $S_{(ij)}$					

The pooled value of  $r_{au}^2$  in (iii) for the numerical example of p. 785 would thus be

$$r_{au}^2 = [(0.73)(0.61) + (0.73)(0.53) + (0.73)(0.44) + (0.61)(0.53) + (0.61)(0.44) + (0.53)(0.44)]$$

$$\div (0.55 + 0.50 + 0.39 + 0.43 + 0.36 + 0.29),$$

$$\therefore r_{au}^2 = 0.79 \quad \text{and} \quad r_{au} = 0.89.$$

As an alternative to this procedure, there is a hit-and-miss method of determining the factor loadings ( $r_{au}$ , etc.), and its rationale is instructive in connexion with analysis of data which do not conform to a single factor pattern. To understand it, we must take stock of relations which subsist between the grand total and column or row totals of the cell entries of the correlation matrix. A 4-test matrix for an exact single factor pattern will suffice to make this clear.

	<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>	
<i>A</i>	$r_{au}^2$	$r_{ab} = r_{au} \cdot r_{bu}$	$r_{ac} = r_{au} \cdot r_{cu}$	$r_{ad} = r_{au} \cdot r_{du}$	
<i>B</i>	$r_{ab} = r_{au} \cdot r_{bu}$	$r_{bu}^2$	$r_{bc} = r_{bu} \cdot r_{cu}$	$r_{bd} = r_{bu} \cdot r_{du}$	
<i>C</i>	$r_{ac} = r_{au} \cdot r_{cu}$	$r_{bc} = r_{bu} \cdot r_{cu}$	$r_{cu}^2$	$r_{cd} = r_{cu} \cdot r_{du}$	
<i>D</i>	$r_{ad} = r_{au} \cdot r_{du}$	$r_{bd} = r_{bu} \cdot r_{du}$	$r_{cd} = r_{cu} \cdot r_{du}$	$r_{du}^2$	
Total	$t_a$	$t_b$	$t_c$	$t_d$	Grand Total $T$

In the foregoing schema  $t_a$ ,  $t_b$ , etc., are column totals and  $T$  is the grand total of the inter-correlations in both dimensions after completing the diagonal entries. We may denote the sum of all the factor loadings ( $a_u = r_{au}$ ,  $b_u = r_{bu}$ , etc.) as

$$s_u = r_{au} + r_{bu} + r_{cu} + r_{du}.$$





TABLE 4

	A	B	C	D	E	F	G	H	J	K
A	$(a_1^2 + a_2^2)$	$a_1b_1 + a_2b_2$	$a_1c_1 + a_2c_2$	$a_1d_1$	$a_1e_1$	$a_1f_1$	$a_1g_1$	$a_1h_1$	$a_1j_1$	$a_1k_1$
B	$a_1b_1 + a_2b_2$	$(b_1^2 + b_2^2)$	$b_1c_1 + b_2c_2$	$b_1d_1$	$b_1e_1$	$b_1f_1$	$b_1g_1$	$b_1h_1$	$b_1j_1$	$b_1k_1$
C	$a_1c_1 + a_2c_2$	$b_1c_1 + b_2c_2$	$(c_1^2 + c_2^2)$	$c_1d_1$	$c_1e_1$	$c_1f_1$	$c_1g_1$	$c_1h_1$	$c_1j_1$	$c_1k_1$
D	$a_1d_1$	$b_1d_1$	$c_1d_1$	$(d_1^2 + d_2^2)$	$d_1e_1 + d_3e_3$	$d_1f_1 + d_3f_3$	$d_1g_1 + d_3g_3$	$d_1h_1$	$d_1j_1$	$d_1k_1$
E	$a_1e_1$	$b_1e_1$	$c_1e_1$	$d_1e_1 + d_3e_3$	$(e_1^2 + e_2^2)$	$e_1f_1 + e_3f_3$	$e_1g_1 + e_3g_3$	$e_1h_1$	$e_1j_1$	$e_1k_1$
F	$a_1f_1$	$b_1f_1$	$c_1f_1$	$d_1f_1 + d_3f_3$	$e_1f_1 + e_3f_3$	$(f_1^2 + f_3^2)$	$f_1g_1 + f_3g_3$	$f_1h_1$	$f_1j_1$	$f_1k_1$
G	$a_1g_1$	$b_1g_1$	$c_1g_1$	$d_1g_1 + d_3g_3$	$e_1g_1 + e_3g_3$	$f_1g_1 + f_3g_3$	$(g_1^2 + g_3^2)$	$g_1h_1$	$g_1j_1$	$g_1k_1$
H	$a_1h_1$	$b_1h_1$	$c_1h_1$	$d_1h_1$	$e_1h_1$	$f_1h_1$	$g_1h_1$	$(h_1^2 + h_3^2)$	$h_1j_1 + h_3j_3$	$h_1k_1 + h_3k_3$
J	$a_1j_1$	$b_1j_1$	$c_1j_1$	$d_1j_1$	$e_1j_1$	$f_1j_1$	$g_1j_1$	$h_1j_1 + h_3j_3$	$(j_1^2 + j_3^2)$	$j_1k_1 + j_3k_3$
K	$a_1k_1$	$b_1k_1$	$c_1k_1$	$d_1k_1$	$e_1k_1$	$f_1k_1$	$g_1k_1$	$h_1k_1 + h_3k_3$	$j_1k_1 + j_3k_3$	$(k_1^2 + k_3^2)$

TABLE 5

$a_2^2$	$a_2b_2$	$a_2c_2$	0	0	0	0	0	0	0	0
$a_2b_2$	$b_2^2$	$b_2c_2$	0	0	0	0	0	0	0	0
$a_2c_2$	$b_2c_2$	$c_2^2$	0	0	0	0	0	0	0	0
0	0	0	$d_3^2$	$d_3e_3$	$d_3f_3$	$d_3g_3$	0	0	0	0
0	0	0	$d_3e_3$	$e_3^2$	$e_3f_3$	$e_3g_3$	0	0	0	0
0	0	0	$d_3f_3$	$e_3f_3$	$f_3^2$	$f_3g_3$	0	0	0	0
0	0	0	$d_3g_3$	$e_3g_3$	$f_3g_3$	$g_3^2$	0	0	0	0
0	0	0	0	0	0	0	$k_3^2$	$h_4j_4$	$h_4k_4$	0
0	0	0	0	0	0	0	$h_4j_4$	$j_4^2$	$j_4k_4$	0
0	0	0	0	0	0	0	$h_4k_4$	$j_4k_4$	$k_4^2$	0



- (a) all the tests are correlated in virtue of a *single* factor common to all ;  
 (b) members of the same group are more highly correlated in virtue of a factor common to and peculiar to the group.

The reader should now find it easy to translate these postulates in terms of the Umpire Bonus Model. Though a satisfactory vindication of the bi-factor pattern by the procedure we shall now examine calls for at least 12 tests, it will suffice for illustrative purposes, if we formulate a schema for *standard* scores of 10 tests involving 3 groups after correction w.r.t. reliability as follows :

$$\begin{aligned} z_a &= a_1 \cdot z_1 + a_2 \cdot z_2 + a_s \cdot z_{sa} \\ z_b &= b_1 \cdot z_1 + b_2 \cdot z_2 + b_s \cdot z_{sb} \\ z_c &= c_1 \cdot z_1 + c_2 \cdot z_2 + c_s \cdot z_{sc} \\ z_d &= d_1 \cdot z_1 + d_3 \cdot z_3 + d_s \cdot z_{sd} \\ z_e &= e_1 \cdot z_1 + e_3 \cdot z_3 + e_s \cdot z_{se} \\ z_f &= f_1 \cdot z_1 + f_3 \cdot z_3 + f_s \cdot z_{sf} \\ z_g &= g_1 \cdot z_1 + g_3 \cdot z_3 + g_s \cdot z_{sg} \\ z_h &= h_1 \cdot z_1 + h_4 \cdot z_4 + h_s \cdot z_{sh} \\ z_j &= j_1 \cdot z_1 + j_4 \cdot z_4 + j_s \cdot z_{sj} \\ z_k &= k_1 \cdot z_1 + k_4 \cdot z_4 + k_s \cdot z_{sk} \end{aligned}$$

For such a schema the correlation matrix is as in Table 4.

Provided that there are at least three groups as in Table 4, in which they are (*ABC*), (*DEFG*), (*HJK*), we can form at least one triad involving only intercorrelations of members of different groups, e.g. :

$$\frac{r_{ad} \cdot r_{ak}}{r_{dk}} = \frac{a_1^2 \cdot d_1 \cdot k_1}{d_1 \cdot k_1} = a_1^2.$$

If each group contains at least 2 members we can test the result, since we can then extract at least one other equivalent triad, e.g. :

$$\frac{r_{ae} \cdot r_{aj}}{r_{ej}} = \frac{a_1^2 \cdot e_1 \cdot j_1}{e_1 \cdot j_1} = a_1^2.$$

Having provisionally vindicated the possibility of extracting values of  $a_1$ ,  $b_1$ , etc., with good agreement between the several estimates of each, we are now in a position to make a matrix exhibiting what the approximate values of the intercorrelations would be, if there were no group factors, i.e. if they arose entirely from the common factor. The first, fourth and last row would then read in conformity with the product rule for the single factor pattern :

$a_1^2$	$a_1 b_1$	$a_1 c_1$	$a_1 d_1$	$a_1 e_1$	$a_1 f_1$	$a_1 g_1$	$a_1 h_1$	$a_1 j_1$	$a_1 k_1$
$a_1 d_1$	$b_1 d_1$	$c_1 d_1$	$d_1^2$	$d_1 e_1$	$d_1 f_1$	$d_1 g_1$	$d_1 h_1$	$d_1 j_1$	$d_1 k_1$
$a_1 k_1$	$b_1 k_1$	$c_1 k_1$	$d_1 k_1$	$e_1 k_1$	$f_1 k_1$	$g_1 k_1$	$h_1 k_1$	$j_1 k_1$	$k_1^2$

On subtracting cell entries of this matrix from the corresponding one of Table 1 we now have a *matrix of residuals* shown in Table 5.

Thus the entries for the between-group correlations in our matrix of residuals offer us an additional check on our procedure, *viz.* none of these residuals should differ significantly from zero. If the result confirms our supposition, we may proceed to assign the values of the group factors  $a_2, e_3, k_4$ , etc., by the method of triad summation implicit in (iii) of 18.07. With appropriate modification the same procedure is adaptable to the extraction of pooled values of the common factor, i.e. to extract a principal factor loading  $a_1$  in group I we form : (a) the numerator by summing all products in the  $A$  row between cell entries of the original matrix after excluding those which belong to the same group as  $A$  itself ; (b) the denominator by adding from one side of the diagonal involving all intercorrelations two members of *different* groups other than the  $A$ -group.

An exacting vindication of the bi-factor set-up will demand that the group residuals are consistent with a single factor patter, i.e. within a group tetrad differences vanish and the triad ratios are consistent. The extraction of the residual matrix by the procedure outlined above itself presupposes that there are at least three group factors, and we have no check on the first step unless each group contains at least 2 members. To validate the last stage we require at least 4 members in each group, since three would provide only a single estimate of any one factor loading.

### 18.09 HIGHER FACTOR PATTERNS

In the foregoing treatment of the bi-factor pattern we postulate a single common factor and 3 group factors. We then have three groups of tests each with two common factors. We shall now ask : is there any procedure by which we could validate our assumptions, if the data suggest the existence of only 2 group factors ? If so, we have 2 groups of tests, with 2 and only 2 factors common to members of the same group. This feature of the bi-factor set-up prompts us to ask : is there any criterion analogous to the tetrad principle to define a set of intercorrelations involving the same two sources of variation and no other ?

An answer to this question is obtainable by recourse to elementary algebra as for the derivation of the tetrad equations in 18.05 ; but the procedure is laborious. The advantage of side-stepping the labour entailed by use of grid algebra in connexion with this problem and *a fortiori* in connexion with the exploration of more complex factor patterns explains the prominent role which matrix algebra plays in expositions of factor analysis, and hence also the prevalent jargon of factor-space and other geometrical metaphors suggested by matrix operations ; but the truth is that the logical assumptions underlying factor analysis are explicable without reliance upon it and to all except the mathematically proficient easier to grasp in the language of the more elementary mathematics employed in foregoing sections of this chapter. Factor analysis presupposes a system of score components expressible as simultaneous linear equations of the type we have met with in connexion with the Umpire Bonus Model. If there are only 2 or 3 relevant variables, recourse to determinants has little value as a labour saving device. There has therefore been no good reason for making the assumptions implicit in factor analysis stepping the labour entailed by use of grid algebra in connexion with this problem and less accessible to readers not at home with determinants and/or matrices by invoking their aid in what has gone before. At this stage, we do so merely because the algebraic treatment of the problem raised above is otherwise very laborious. There is no indispensable tie-up between the logic of factor analysis and the matrix algebra which most expositions of the procedure invoke at an introductory stage as a prerequisite to understanding it.

Before proceeding further let us recall the structure of a tetrad equation, e.g.

$$r_{ab} \cdot r_{kq} - r_{ak} \cdot r_{bq} = 0.$$



In this expression  $r_{ab}$  and  $r_{ak}$  are cell entries of the same ( $A$ -) row,  $r_{ak}$  and  $r_{kq}$  are cell entries of the same ( $K$ -) column,  $r_{ab}$  and  $r_{bq}$  occur in the same ( $B$ -) column, while  $r_{kq}$  and  $r_{bq}$  occur in the same ( $Q$ -) row. Thus the 4 relevant entities occur as cell entries at the apices of a rectangular segment of the matrix and as such constitute a *minor* of the first order, *viz.* :

$$\begin{vmatrix} r_{ab} & r_{ak} \\ r_{bq} & r_{kq} \end{vmatrix} = r_{ab} \cdot r_{kq} - r_{ak} \cdot r_{bq}.$$

We might therefore express the tetrad rule derived in 18.04 by saying that one common factor and only one suffices to specify a matrix of intercorrelations if *every determinant minor of the first order vanishes*.

This suggests a more general rule due to Thurstone. For our purpose it will suffice to demonstrate it for a group of tests each pair of which involves 2 common factors and two only. The rule then states that all *2nd order* minors of the correlation matrix must vanish. Below we set out such a minor: (*a*) in terms of the cell entries; (*b*) in terms of the loading factors prescribed in 18.06 and 18.07.

$$\begin{vmatrix} r_{be} & r_{bk} & r_{bm} \\ r_{ge} & r_{gk} & r_{gm} \\ r_{je} & r_{jk} & r_{jm} \end{vmatrix} = \Delta_3 = \begin{vmatrix} b_1e_1 + b_2e_2 & b_1k_1 + b_2k_2 & b_1m_1 + b_2m_2 \\ g_1e_1 + g_2e_2 & g_1k_1 + g_2k_2 & g_1m_1 + g_2m_2 \\ j_1e_1 + j_2e_2 & j_1k_1 + j_2k_2 & j_1m_1 + j_2m_2 \end{vmatrix}$$

We may reduce the determinant on the right as follows :

$$\begin{aligned} \frac{\Delta_3}{e_1k_1m_1} &= \begin{vmatrix} b_1 + b_2 \frac{e_2}{e_1} & b_1 + b_2 \frac{k_2}{k_1} & b_1 + b_2 \frac{m_2}{m_1} \\ g_1 + g_2 \frac{e_2}{e_1} & g_1 + g_2 \frac{k_2}{k_1} & g_1 + g_2 \frac{m_2}{m_1} \\ j_1 + j_2 \frac{e_2}{e_1} & j_1 + j_2 \frac{k_2}{k_1} & j_1 + j_2 \frac{m_2}{m_1} \end{vmatrix} \\ &= \begin{vmatrix} b_2 \left( \frac{e_2}{e_1} - \frac{m_2}{m_1} \right) & b_2 \left( \frac{k_2}{k_1} - \frac{m_2}{m_1} \right) & b_1 + b_2 \frac{m_2}{m_1} \\ g_2 \left( \frac{e_2}{e_1} - \frac{m_2}{m_1} \right) & g_2 \left( \frac{k_2}{k_1} - \frac{m_2}{m_1} \right) & g_1 + g_2 \frac{m_2}{m_1} \\ j_2 \left( \frac{e_2}{e_1} - \frac{m_2}{m_1} \right) & j_2 \left( \frac{k_2}{k_1} - \frac{m_2}{m_1} \right) & j_1 + j_2 \frac{m_2}{m_1} \end{vmatrix} \\ &= \left( \frac{e_2}{e_1} - \frac{m_2}{m_1} \right) \left( \frac{k_2}{k_1} - \frac{m_2}{m_1} \right) \begin{vmatrix} b_2 & b_2 & b_1 + b_2 \frac{m_2}{m_1} \\ g_2 & g_2 & g_1 + g_2 \frac{m_2}{m_1} \\ j_2 & j_2 & j_1 + j_2 \frac{m_2}{m_1} \end{vmatrix} \end{aligned}$$

The determinant in the last expression has two identical rows, whence its numerical value is zero and

$$\Delta_3 = 0.$$

This completes the proof that the 2nd order minors vanish, if all the intercorrelations are referable to the same 2 common factors. The converse assertion is demonstrable. As an exercise, the student may paint in the values of  $r_{be}$ , etc., for a 3-factor set-up, e.g.

$$r_{be} = b_1e_1 + b_2e_2 + b_3e_3.$$

We then find that  $\Delta_3$  vanishes only if one of the factors is algebraically redundant. In short, the full statement of Thurstone's rule illustrates one of the uses of matrix algebra in the theory of equations, i.e. to prescribe the redundancy or otherwise of one or more variables. Geometrical terms employed in the theory of equations are suggestive to the mathematician who is already at home with them; but this is not essential to an appreciation of what a particular factor pattern postulates. We could, of course, express Thurstone's rule for the 2-factor pattern in the form it takes when we expand the 2nd order determinant, *viz.* :

$$r_{be}(r_{gk}r_{jm} - r_{gm}r_{jk}) - r_{bk}(r_{ge}r_{jm} - r_{gm}r_{je}) + r_{bm}(r_{ge}r_{jk} - r_{gk}r_{je}) = 0.$$

The student who wishes to check this by lower school certificate algebra may do so. The attempt will at least dispose of any lingering doubts concerning the advantages of a little familiarity with the use of determinants. It will also be a wholesome demonstration of the irrelevance of portentous excursions into matrix algebra as a preliminary to understanding the logical assumptions which factor analysis prescribes.

The rule we have last examined raises the question, how many tests in a group suffice to identify a 2-factor pattern with due regard to the fact that our observational data do not furnish cell entries for the diagonal of the correlation matrix? For a single factor pattern we need 4 tests to provide a complete first order minor. This is evident, if we set out as below a 3-test matrix on the left and a 4-test matrix on the right :

$$\begin{bmatrix} - & r_{ab} & r_{ac} \\ r_{ab} & - & r_{bc} \\ r_{ac} & r_{bc} & - \end{bmatrix} \quad \begin{bmatrix} - & r_{ab} & r_{ac} & r_{ad} \\ r_{ab} & - & r_{bc} & r_{bd} \\ r_{ac} & r_{bc} & - & r_{cd} \\ r_{ad} & r_{bd} & r_{cd} & - \end{bmatrix}$$

Evidently, we cannot get a 2nd order minor from the 4-test matrix. Nor can we do so with 5 tests. We need at least 6, as below.

$$\begin{array}{cccccc} - & r_{ab} & r_{ac} & r_{ad} & r_{ae} & r_{af} \\ r_{ab} & - & r_{bc} & r_{bd} & r_{be} & r_{bf} \\ r_{ac} & r_{bc} & - & r_{cd} & r_{ce} & r_{cf} \\ r_{ad} & r_{bd} & r_{cd} & - & r_{de} & r_{df} \\ r_{ae} & r_{be} & r_{ce} & r_{de} & - & r_{ef} \\ r_{af} & r_{bf} & r_{cf} & r_{df} & r_{ef} & - \end{array}$$

Theoretically, then, we can check up on the requirements of a bi-factor pattern involving one common and 2 group factors if each of the two groups contains at least 6 members or twelve in all. By applying Thurstone's rule to each group and to the entire matrix we can establish the conclusions: (a) 2 factors suffice to explain correlations within the group since second order minors vanish; (b) one factor suffices to explain correlations between groups since inter-group tetrads vanish, e.g. in the schema of 18.04,  $r_{ad} \cdot r_{cf} = r_{af} \cdot r_{cd}$ .

It remains to ask: can we assign to the common factor values from which we can reconstruct by subtraction (as in 18.08) a matrix of residuals having values consonant with the requirements that: (a) inter-group residuals are zero; (b) intra-group residuals are hierarchical? To clarify the situation the student may set out in full the schema of *inter-group*



correlations for the single common factor components of the matrix for a battery of 12 tests,  $A-F$  being one group and  $G-M$  the other ; but it will suffice for illustrative purposes if we take 3 of one group and 3 of the other as below :

	$G$	$H$	$J$	Total	
	Factor	$g_1$	$h_1$	$j_1$	$s_g$
$A$	$a_1$	$r_{ag} = a_1 g_1$	$r_{ah} = a_1 h_1$	$r_{aj} = a_1 j_1$	$t_a = a_1 s_g$
$B$	$b_1$	$r_{bg} = b_1 g_1$	$r_{bh} = b_1 h_1$	$r_{bj} = b_1 j_1$	$t_b = b_1 s_g$
$C$	$c_1$	$r_{cg} = c_1 g_1$	$r_{ch} = c_1 h_1$	$r_{cj} = c_1 j_1$	$t_c = c_1 s_g$
Total	$s_a$	$t_g = g_1 s_a$	$t_h = h_1 s_a$	$t_j = j_1 s_a$	$T = s_a s_g$

Evidently the hierarchical principle obtains for all inter-group correlations, but the reader who is familiar with the theory of equations will find that such a schema yields too few independent equations to admit of a unique solution for the factor loadings. Since one exception suffices to dismiss the possibility, it is instructive to exhibit two complete solutions as below :

		Total			
	Factor	(0.3)	(0.4)	(0.5)	(.12)
	(0.2)	0.06	0.08	0.10	0.24
	(0.5)	0.15	0.20	0.25	0.60
	(0.1)	0.03	0.04	0.05	0.12
Total	(0.8)	0.24	0.32	0.40	0.96

		Total			
	Factor	(0.24)	(0.32)	(0.40)	(0.96)
	(0.250)	0.06	0.08	0.10	0.24
	(0.625)	0.15	0.20	0.25	0.60
	(0.125)	0.03	0.04	0.05	0.12
Total	(1.0)	0.24	0.32	0.40	0.96

The restrictive relations in the evaluation are all included in the system of equations exhibited in the row and column totals of the schema

$$g_1 = \frac{t_g}{s_a}; \quad h_1 = \frac{t_h}{s_a}; \quad j_1 = \frac{t_j}{s_a};$$

$$a_1 = \frac{t_a}{s_g}; \quad b_1 = \frac{t_b}{s_g}; \quad c_1 = \frac{t_c}{s_g};$$

$$T = s_a \cdot s_g.$$

By starting with an arbitrary value of  $s_a$ , the reader may easily check that any set of values consistent with the above works. For example,  $s_a = 3$  implies that  $s_g = 0.32$  to give  $T = 0.96$ . We then have

$$g_1 = \frac{0.24}{3} = 0.08; \quad a_1 = \frac{0.24}{0.32} = 0.75; \quad r_{ag} = (0.08)(0.75) = 0.06, \text{ etc.}$$

Now any set of common factor loadings such as the above would lead to the same residual matrix from which to determine our group factor loadings. The situation is therefore this: we can recognise what the factor pattern is; but we cannot assign unique values to the factor loadings.

This lack of uniqueness is the pivot of a controversy which involves two issues. On the *analysis*, if the factor loadings they prescribe are arbitrary. On the other hand, it may be instructive to recognise the existence of a factor pattern in the absence of a numerical specification of the loadings; and we have seen that this may be possible to accomplish when no unique numerical solution is realisable. If so, it is important to be clear about what we mean by recognising the pattern.

Against the background of the Umpire Bonus Model, the reader who cares to pursue the topic will find the following hints helpful. Of two essentially different procedures subsumed by the term *multiple* factor analysis, that of Thurstone admits the possibility that: (a) a score referable to a particular test may contain a *specific* factor component; (b) the number of non-specific factors may be as great as the number of different pairs of tests in the battery. In the idiom of the Umpire Bonus Model this means that each player's total score contains some multiple—not excluding zero as a possible multiplier—of his individual score together with some multiple—not excluding zero as a possible multiplier—of each of different umpire contributions equal in number to the number of pairs of players. The initial formulation definitive of score value components thus leads to a system of fewer equations than variables. The solution sought by an iterative procedure is the most economical in the sense that it seeks to interpret consistently any inter-test correlation in terms of the least number of score components.

The assumptions implicit in Hotelling's procedure are on all fours with the rules of the model 3-wheel game prescribed at the end of 12.07. There the player is passive, each player's score being made up of contributions from the same number of umpires not exceeding the number of different pairs of players in the initial formulation of the Hotelling set-up. The number of basic equations definitive of score components cannot therefore exceed the number of variables, a circumstance which confers on the game an aspect of greater algebraic propriety than that of Thurstone. Again, however, we must invoke a quite arbitrary axiom of economy in the search for a satisfactory selection of one among many consistent sets of factor loadings; and the prescribed procedure will lead, as pointed out by Godfrey Thomson, to different factor loadings if we add a new test—and hence the possibility of  $n$  new factors to an  $n$ -fold test battery.

Either method involves an issue which is outside the scope of mathematics as such. Thurstone himself faces it frankly when he appeals to William of Occam's principle: *entia non multiplicandur praeter necessitatem*. According to his view the most economical factor loadings are the best because economy of hypothesis is a canon of scientific method. This plea is open to criticism at more than one level. The use of the term economy in scientific enquiry is not wholly unequivocal; and what G. P. Meredith calls the *epistemic* status of the canon is itself debatable. As the writer sees it, a methodical scientific worker will rightly choose to investigate first the simpler of two hypotheses to forestall unnecessary waste of effort, if it stands the test of experience equally well. So interpreted the Occam principle embodies a wise code of procedure in the process of discovery; but embodies no rational prescription for deciding whether one or the other hypothesis is true or false without appeal to the higher tribunal of the *experimentum crucis*.



All such procedures referred to as factor analysis presume a strictly additive relation between the score components and zero covariance between any two of them. In the absence of confirmatory evidence, this assumption, which is also inherent in the construction of the balance sheet of the analysis of variance, is highly arbitrary and sometimes grossly inappropriate. In Chapter 13, we have seen how the replication criterion may give us the opportunity of confirming the additive postulate when the end in view is a balance sheet of variance, and we have seen reason to believe that the same principle is both a necessary and sufficient condition of the validity of Thurstone's rule and the tetrad criterion as a special case of it. Just as it is all too common practice to overlook the importance of the replication criterion, it is all too common to execute elaborate computations to extract factor loadings when there is : (a) no other clear-cut factor pattern to validate the initial assumption of the additivity of the score components and the twin postulate of zero covariance ; (b) no possibility of arriving at a solution preferable to others equally consistent with the data.

## CHAPTER 19

# SAMPLING IN A FINITE UNIVERSE AND MANIFOLD CLASSIFICATION

### 19.01 THE HYPERGEOMETRIC DISTRIBUTION

We have already (Chapter 2, Vol. I) obtained the exact expression for sampling without replacement in the domain of binary taxonomic scoring, *viz.* a distribution defined by successive terms of the binomial in factorial powers :

$$\frac{(f+s)^{(r)}}{n^{(r)}} = \frac{(nq + np)^{(r)}}{n^{(r)}}.$$

If the sampling fraction ( $F$ ) defined by the ratio  $nF = r$  is small and  $n$  is very great, we have seen (Chapter 3, Vol. I) that we may regard the distribution as approximately normal ; but we have not as yet explored the possibility of finding a satisfactory fitting curve when  $n$  is great if  $F$  itself is not a small fraction. We shall now do so by recourse to the method of moments. For reasons which we have seen in 14.04, it will be convenient to derive first a general expression for the *factorial* moments, defined as

$$\mu_{(k)} = \frac{1}{n^{(r)}} \sum_{x=0}^{x=r} x^{(k)} \frac{r!}{x!(r-x)!} s^{(x)} f^{(r-x)} \quad . \quad . \quad . \quad . \quad (i)$$

In this expression

$$r! = r^{(k)}(r-k)!; \quad s^{(x)} = s^{(k)}(s-k)^{(x-k)};$$

$$\frac{x^{(k)}}{x!} = \frac{1}{(x-k)!}$$

$$\therefore \mu_{(k)} = \frac{r^{(k)} s^{(k)}}{n^{(r)}} \sum_{x=0}^{x=r} \frac{(r-k)!}{(x-k)! (r-x)!} (s-k)^{(x-k)} f^{(r-x)} \quad . \quad . \quad . \quad . \quad (ii)$$

If we write  $(r-k) = a$  and  $(x-k) = b$  :

$$\sum_{x=0}^{x=r} \frac{(r-k)!}{(x-k)! (r-x)!} (s-k)^{(x-k)} f^{(r-x)} = \sum_{b=-k}^{b=a} \frac{a!}{b! (a-b)!} (s-k)^{(b)} f^{(a-b)}.$$

When  $b < 0$ , the reciprocal of  $b!$  is zero, whence

$$\begin{aligned} \sum_{x=0}^{x=r} \frac{(r-k)!}{(x-k)! (r-x)!} (s-k)^{(x-k)} f^{(r-x)} &= \sum_{b=0}^{b=a} \frac{a!}{b! (a-b)!} (s-k)^{(b)} f^{(a-b)} \\ &= (s+f-k)^{(a)} \\ &= (n-k)^{(r-k)}. \end{aligned}$$

Whence by substitution in (ii)

$$\mu_{(k)} = \frac{r^{(k)} s^{(k)} (n-k)^{(r-k)}}{n^{(r)}} \quad . \quad . \quad . \quad . \quad . \quad (iii)$$





Whence we have

$$m_3 = \frac{r^{(3)}s^{(3)}}{n^{(3)}} - \frac{3r^{(2)}s^{(2)}(rs - n)}{n \cdot n^{(2)}} + \frac{rs(2rs - n)(rs - n)}{n^3},$$

$$\therefore m_3 = \frac{r pq(q - p)(n - r)(n - 2r)}{(n - 1)(n - 2)} \quad \text{. . . . . (xii)}$$

On the assumption that the universe is large, so that  $(n - 2) \simeq n$ , we therefore have

$$m_3 \simeq r pq(q - p)(1 - F)(1 - 2F) \quad \text{. . . . . (xiii)}$$

In the same way we obtain

$$m_4 = r pq(n - r)[n(n + 1) - 6r(n - r) + 3pq\{n^2(r - 2) - nr^2 + 6r(n - r)\}] \div (n - 1)^{(3)} \quad \text{(xiv)}$$

$$\therefore m_4 \simeq r pq(1 - F)[1 - 6F(1 - F)(1 - 3pq) + 3pq(r - 2 - nF^2)] \quad \text{. . . . . (xv)}$$

All the expressions for the moments cited above reduce to those of the replacement distribution, when  $F = 0$  as must be true of a finite  $r$ -fold sample from an infinite ( $n = \infty$ ) universe. The appearance of the factor  $(1 - 2F)$  in (xiii) shows that the third moment vanishes when the sampling fraction is  $\frac{1}{2}$ . In fact, all odd moments then vanish, as we can see from the following considerations. Let us consider the frequencies ( $y_1$  and  $y_2$ ) referable to score values  $x_1$  and  $x_2$  equidistant from the mean on either side of it, so that we put  $x_1 = (M - a)$  and  $x_2 = (M + a)$ . If  $F = \frac{1}{2}$ , we have  $(n - s) = f = (2r - s)$ , and

$$\frac{y_1}{y_2} = \frac{s^{(x_1)}(2r - s)^{(r - x_1)}}{x_1(r - x_1)!} \cdot \frac{x_2!(r - x_2)!}{s^{(x_2)}(2r - s)^{(r - x_2)}}$$

$$= \frac{s^{(M - a)}(2r - s)^{(r - M + a)}}{(M - a)!(r - M + a)!} \cdot \frac{(M + a)!(r - M - a)!}{s^{(M + a)}(2r - s)^{(r - M - a)}}.$$

On substituting in accordance with the formula  $c^{(x)} \cdot (c - x)! = c!$ , this reduces to

$$\frac{y_1}{y_2} = \frac{(M + a)!(s - M - a)!(r - s + M + a)!(r - M - a)!}{(M - a)!(s - M + a)!(r - s + M - a)!(r - M + a)!}.$$

Since  $M = rp$  and  $s = np$ , the condition  $F = \frac{1}{2}$  implies that  $M = \frac{1}{2}s$ , whence

$$\frac{y_1}{y_2} = \frac{(\frac{1}{2}s + a)!(\frac{1}{2}s - a)!(r - \frac{1}{2}s + a)!(r - \frac{1}{2}s - a)!}{(\frac{1}{2}s - a)!(\frac{1}{2}s + a)!(r - \frac{1}{2}s - a)!(r - \frac{1}{2}s + a)!}$$

$$\therefore \frac{y_1}{y_2} = 1.$$

Thus the frequencies of any two score values equidistant from the mean are identical when  $F = \frac{1}{2}$  and the  $r$ -fold sample distribution is symmetrical whether  $p$  is equal to  $q$ , greater than  $q$  or less than  $q$ .

Thus  $\beta_1$  vanishes if  $F = \frac{1}{2}$ , just as it also vanishes if  $p = q$ , i.e.  $s = \frac{1}{2}n$ . Subject to the same condition, (ix) and (xv) reduce to

$$m_2 = \frac{1}{2}rpq;$$

$$m_4 = \frac{1}{2}rpq \left[ 1 - \frac{3(1 - 3pq)}{2} + \frac{3(r - 4)pq}{2} \right] r$$

$$\therefore \beta_2 = 3 - \frac{3pq + 1}{rpq} \quad \text{. . . . . (xvi)}$$



This has the same form as the second Pearson coefficient defined by (xx) in 15.04 for the symmetrical  $B(j, k)$  variate of restricted range (Type II). This we have seen to be

$$\beta_2 = 3 - \frac{6}{2j + 3}.$$

This is equivalent to the above when

$$j = \frac{6pq(r - 1) - 3(pq + 1)}{6pq + 2}.$$

To satisfy the requirements of Type II, it is also necessary to show that

$$\mu_1 = \frac{a}{2} \quad \text{and} \quad m_2 = \frac{a^2}{4(2j + 1)},$$

$$\therefore j = \frac{\mu_1^2}{2m_2} - \frac{1}{2}.$$

In our expression  $\mu_1 = rp$  and  $2m_2 = rpq$  when  $F = \frac{1}{2}$ , so that the above also implies the relation

$$j = \frac{rp}{q} - \frac{1}{2}.$$

When  $p = \frac{1}{2} = q$ , the distribution is again symmetrical regardless of the size of the sampling fraction, so that

$$\beta_1 = 0 \quad \text{and} \quad \beta_2 = 3 - \frac{6F(1 - F) + 2}{r(1 - F)}.$$

This again is consistent with one of the requirements of the Type II distribution if we write

$$j = \frac{3(1 - F)(r - 3F) - 3}{6F(1 - F) + 2}.$$

More generally, when neither  $p$  nor  $F$  is equal to  $\frac{1}{2}$ , the form of the  $\beta$  coefficient may conform with Type I requirements. The determination of the constants is laborious, and the student who wishes to pursue the topic may consult Pearson's tables of the Incomplete Beta Function.

## 19.02 MOMENTS OF A SCORE-SUM DISTRIBUTION

The results derived in 19.01 are obtainable from more general expressions for the moments of a score-sum distribution in the domain of representative scoring without replacement. To establish them, a digression is necessary. It is the peculiarity of binary taxonomic scoring that the algebraic form of the unit sample distribution is inherent in the statement of the problem. Representative scoring—except in the limiting case of a 2-class system—raises a new problem. In the foregoing section, we have determined the moments of the non-replacement distribution of samples from a binary universe to obtain an approximate expression for the distribution of the  $r$ -fold sample. In what follows we may remind ourselves that a continuous distribution can give a satisfactory description of a universe only if the number of score classes is very large, and hence that we may usually disregard the consequences of non-replacement. We therefore assume that the unit sample distribution is discrete. In practice, we may likewise assume the unlikelihood that our manifold universe of scores closely conforms to any known *discrete* sampling distribution such as the rectangular or the distribution defined by successive terms of a binomial or of a Vandemonde expansion. It will suffice to postulate that we have empirical sources of

information from which to determine the moments of the unit sample distribution, and on that understanding we shall derive an expression for the moments of the  $r$ -fold distribution without recourse to other data.

Subject to the replacement condition or to the postulate that the size of the universe and/or sampling fraction permits us to disregard it, we have obtained expressions of this sort in 14.05 by the method of iteration. We shall now show how it is possible to derive them by a different procedure which is adaptable to situations in which : (a) there is no replacement ; (b) extraction of the sample materially changes the composition of the parent universe.

We customarily define the representative score of a sample by the mean of the constituent individual scores, i.e. by the quotient of their sum and the sample size. The latter being constant for a sample of given size is immaterial to a specification of the distribution, since all the parameters of the distribution of the mean are obtainable from those of the sum ( $S_r$ ) of the  $r$ -fold sample by a scalar change involving  $r$  alone. We can regard the  $r$ -fold sample as  $r$  successively extracted unit samples of score  $x_u$ , so that

$$S_r = x_1 + x_2 + x_3 \quad . \quad . \quad . \quad x_r = \sum_{u=1}^{u=r} x_u.$$

By definition the  $k$ th zero moment of the score-sum is the expected value of  $S_r^k$  ; and we may write this as

$$\mu_k(S_r) = E(x_1 + x_2 + x_3 \dots x_r)^k.$$

An examination of the moments of the 3-fold sample distribution brings into focus what we need to know in order to evaluate an expression such as the above. Thus we have

$$\mu_2(S_3) = E(x_1 + x_2 + x_3)^2 = E(x_1^2) + E(x_2^2) + E(x_3^2) + 2E(x_1 \cdot x_2) + 2E(x_1 \cdot x_3) + 2E(x_2 \cdot x_3).$$

Now the subscripts we attach to  $x$  in this expression refer to the order in which we extract the unit samples, regardless of the numerical value any individual unit score  $x_u$  or  $x_v$  may have. This expression therefore contains terms of two sorts : (a) squares of unit scores ; (b) products of unit scores whose numerical values may be the same or different. If we replace each item to which we attach a score before drawing another it is evident that : (i) the numerical value of the square unit score does not depend on the value of the subscript  $u$  or  $v$  ; (ii) the numerical value of the product of different unit scores is likewise independent of the particular values assigned to  $u$  and to  $v$ . Whence we may write the last expression in the form

$$\mu_2(S_3) = 3E(x_u^2) + 6E(x_u \cdot x_v).$$

Similarly we may write

$$\begin{aligned} \mu_3(S_3) &= E(x_1^3) + E(x_2^3) + E(x_3^3) + 3E(x_1^2 \cdot x_2) \\ &\quad + 3E(x_1^2 \cdot x_3) + 3E(x_2^2 \cdot x_1) + 3E(x_2^2 \cdot x_3) \\ &\quad + 3E(x_3^2 \cdot x_1) + 3E(x_3^2 \cdot x_2) + 6E(x_1 \cdot x_2 \cdot x_3), \\ \therefore \mu_3(S_3) &= 3E(x_u^3) + 18E(x_u^2 \cdot x_v) + 6E(x_u \cdot x_v \cdot x_w). \end{aligned}$$

In these expressions  $E(x_u^2)$  and  $E(x_u^3)$  are respectively the expected values of the square and the cube of the unit sample score, i.e. of the second and third moments of the unit sample distribution ; and we may write these as  $\mu_2$  and  $\mu_3$  respectively. We may speak of a  $k$ -fold *co-moment* as the expected value of the product of the unit scores of a  $k$ -fold sub-sample, and write

$$\begin{aligned} E(x_u \cdot x_v) &= \mu_{1.1} ; \quad E(x_u^2 \cdot x_v) = \mu_{2.1} = E(x_u \cdot x_v^2) ; \\ E(x_u \cdot x_v \cdot x_w) &= \mu_{1.1.1}, \text{ etc.} \end{aligned}$$



In this symbolism, and subject to the replacement condition unless we can subsequently show its irrelevance,

$$\mu_2(S_3) = 3\mu_2 + 6\mu_{1.1};$$

$$\mu_3(S_3) = 3\mu_3 + 18\mu_{2.1} + 6\mu_{1.1.1}.$$

When replacement does occur we may write

$$E(x_u^j \cdot x_v^k) = E(x_u^j)E(x_v^k);$$

$$E(x_u^h \cdot x_v^i \cdot x_w^j) = E(x_u^h)E(x_v^i)E(x_w^j),$$

$$\therefore \mu_{j.k} = \mu_j \cdot \mu_k \quad \text{and} \quad \mu_{h.i.j} = \mu_h \cdot \mu_i \cdot \mu_j.$$

Hence we may write

$$\mu_2(S_3) = 3\mu_2 + 6\mu_1^2;$$

$$\mu_3(S_3) = 3\mu_3 + 18\mu_2 \cdot \mu_1 + 6\mu_1^3.$$

In the same symbolism, as the reader may check by expanding  $(x_1 + x_2 + x_3)^4$ ,

$$\mu_4(S_3) = 3\mu_4 + 24\mu_3 \cdot \mu_1 + 18\mu_2^2 + 36\mu_2 \cdot \mu_1^2.$$

These considerations suggest the possibility of obtaining general expressions for the zero moments of the  $r$ -fold score-sum of a replacement distribution if we can enumerate terms involving the same set of exponents in the expansion of the expression  $(x_1 + x_2 + x_3 \dots x_m)^k$ . It will clarify the issue if we re-examine the derivation of the multinomial theorem by recourse to the chessboard device.

### 19.03 CHESSBOARD DERIVATION OF THE MULTINOMIAL THEOREM

The chessboard device is at once a replica of the algorithm of multiplication and a means of exhibiting all possible permutations consistent with repetition. Our use of it to derive the binomial sample distribution in Chapter 1 of Vol. I is a particular case of its successive application to exhibit the build up of the multinomial, as below

		<i>a</i>	<i>b</i>	<i>c</i>					
<i>a</i>	<i>aa</i>	<i>ab</i>	<i>ac</i>						
<i>b</i>	<i>ba</i>	<i>bb</i>	<i>bc</i>						
<i>c</i>	<i>ca</i>	<i>cb</i>	<i>cc</i>						

	<i>aa</i>	<i>ab</i>	<i>ac</i>	<i>ba</i>	<i>bb</i>	<i>bc</i>	<i>ca</i>	<i>cb</i>	<i>cc</i>
<i>a</i>	<i>aaa</i>	<i>aab</i>	<i>aac</i>	<i>aba</i>	<i>abb</i>	<i>abc</i>	<i>aca</i>	<i>acb</i>	<i>acc</i>
<i>b</i>	<i>baa</i>	<i>bab</i>	<i>bac</i>	<i>bba</i>	<i>bbb</i>	<i>bbc</i>	<i>bca</i>	<i>bcb</i>	<i>bcc</i>
<i>c</i>	<i>caa</i>	<i>cab</i>	<i>cac</i>	<i>cba</i>	<i>cbb</i>	<i>cbc</i>	<i>cca</i>	<i>ccb</i>	<i>ccc</i>

If we take out all terms in the above with identical factors regardless of order we may classify them as follows

$$\begin{aligned}
 aaa &= a^3; \quad bbb = b^3; \quad ccc = c^3 \\
 aab + aba + baa &= 3a^2b \\
 aac + aca + caa &= 3a^2c \\
 bba + bab + abb &= 3b^2a \\
 bbc + bcb + cbb &= 3b^2c \\
 cca + cac + acc &= 3c^2a \\
 ccb + cbc + bcc &= 3c^2b \\
 abc + acb + bac + bca + cab + cba &= 6abc.
 \end{aligned}$$

When, as is customary, we collect our terms in this way, the appropriate numerical coefficient of each one is implicit in the law of generation inherent in successive application of the chessboard lay-out. We may state it thus. The exponent of each *basic term* ( $a, b, c$ , etc.) is the number of times the latter appears as a factor in the product. In the expansion of  $(a + b + c \dots)^n$  each product will have  $n$  factors in this sense; and the number of identical factors may be 1, 2, 3, ...  $m$  if  $m$  is the number of basic terms. If  $u, v, w$ , etc., signify the exponents of  $a, b, c$ , etc., they therefore respectively represent how many times  $a, b, c$  appear as factors in the product, and their sum is  $n$ . The numerical coefficient of each product with the same build-up regardless of order is simply the number of linear permutations consistent with its build-up; and the number of different linear arrangements consistent with the build-up of a term of the form  $a^u b^v c^w$  is the number of ways in which we can set out in a row  $n$  cards classifiable as three classes respectively composed of  $u, v$ , and  $w$  members. This is given by the familiar formula

$${}^n P_{u \cdot v \cdot w} = \frac{n!}{u! v! w!} \quad \dots \quad (i)$$

Thus we speak of  ${}^n P_{u \cdot v \cdot w}$  so defined as the coefficient of the general term of the multinomial of  $n$ th degree. As we have already seen in Chapter 1 of Vol. I, the sum of all the numerical coefficients so defined is  $m^n$ , as is evident from the identity

$$m^n = (1 + 1 + 1 \dots m \text{ times})^n = \sum_{u=0}^{u=n} \sum_{v=0}^{v=n-u} \dots \frac{n!}{u! v! w! \dots} \quad (ii)$$

If  $a = b = c$ , etc., in the multinomial  $(a + b + c \dots)^n$ , it will be evidently convenient to carry our classification a step further by collecting products of the *same order*, i.e. with identical exponents regardless of the component (and then numerically equivalent) basic terms. For instance we might write out the expansion of  $(a + b + c)^3$  in accordance with the following schema:

$$\begin{aligned}
 a^3 + b^3 + c^3 &= 3h^3; \\
 3a^2b + 3a^2c + 3b^2a + 3b^2c + 3c^2a + 3c^2b &= 18h^2i; \\
 6abc &= 6hij.
 \end{aligned}$$

Similarly, we should write

$$\begin{aligned}
 (a + b + c)^2 &= 3h^2 + 6hi; \\
 (a + b + c + d)^4 &= 4h^4 + 48h^3i + 36h^2i^2 + 144h^2ij + 24hijk.
 \end{aligned}$$

Necessarily, the rule exhibited in (ii) holds good since we have merely collected terms with the common property that they contain the same assemblage of exponents. Thus the sum of the





Whence we derive

$$K_{p, q, r} = \frac{m^{(e)}}{s_1! s_2! s_3! \dots} \cdot \frac{n!}{u! v! w! \dots} \quad (v)$$

In this symbolism

$$K_{1.2.3} = K_{2.1.3} = K_{3.1.2}, \text{ etc.} = \frac{m^{(3)}}{1! 1! 1!} \cdot \frac{6!}{3! 2! 1!},$$

$$K_{2.2.1} = K_{2.1.2} = K_{1.2.2} = \frac{m^{(3)}}{2! 1!} \cdot \frac{5!}{2! 2! 1!},$$

$$K_{3.4.4} = K_{4.3.4} = K_{3.4.4} = \frac{m^{(3)}}{2! 1!} \cdot \frac{11!}{4! 3! 4!},$$

$$K_{3.3} = \frac{m^{(2)}}{2!} \cdot \frac{6!}{3! 3!}.$$

To clarify the meaning of (v) let us consider the expansion of  $(a + b + c)^4$ . We may expand this in terms of the same order as follows:

$$(a + b + c)^4 = K_4 h^4 + K_{3.1} h^3 i + K_{2.2} h^2 i^2 + K_{2.1.1} h^2 i j + K_{1.1.1.1} h i j k.$$

In this expansion  $m = 3$  and  $n = 4$ , the values of  $e$  corresponding to the numerical coefficients being successively 1, 2, 2, 3 and 4. In two products ( $h^2 i^2$  and  $h^2 i j$ ) the same exponent (2 and 1 respectively) occurs twice. Otherwise no exponent occurs more than once. Thus we have

$$\begin{aligned} K_4 &= \frac{3^{(1)}}{1!} \cdot \frac{4!}{4! 0! 0!} = 3, \\ K_{3.1} &= \frac{3^{(2)}}{1! 1!} \cdot \frac{4!}{3! 1! 0!} = 24, \\ K_{2.2} &= \frac{3^{(2)}}{2!} \cdot \frac{4!}{2! 2! 0!} = 18, \\ K_{2.1.1} &= \frac{3^{(3)}}{1! 2!} \cdot \frac{4!}{2! 1! 1!} = 36, \\ K_{1.1.1.1} &= \frac{3^{(4)}}{4!} \cdot \frac{4!}{1! 1! 1! 1!} = 0 \\ \text{Total} & \quad \quad \quad 81 \end{aligned}$$

The total 81 checks, being  $3^4 = m^n$ . More fully, we have

$$\begin{aligned} a^4 + b^4 + c^4 &= 3h^4, \\ 4a^3b + 4a^3c + 4b^3a + 4b^3c + 4c^3a + 4c^3b &= 24h^3i, \\ 6a^2b^2 + 6a^2c^2 + 6b^2c^2 &= 18h^2i^2, \\ 12a^2bc + 12b^2ac + 12c^2ab &= 36h^2ij. \end{aligned}$$

From the foregoing, we may write down the general expressions for the expansion of orders  $u = 2, 3, 4$  as follows:

$$(a + b + c \dots)^2 = m^{(1)}h^2 + m^{(2)}hi \quad (vi)$$

$$(a + b + c \dots)^3 = m^{(1)}h^3 + 3m^{(2)}h^2i + m^{(3)}hij \quad (vii)$$

$$(a + b + c \dots)^4 = m^{(1)}h^4 + 4m^{(2)}h^3i + 3m^{(2)}h^2i^2 + 6m^{(3)}h^2ij + m^{(4)}hijk \quad (viii)$$









Now the coefficients of  $E(x_u^k)$ ,  $E(x_u^k \cdot x_v^m)$ , etc., in the foregoing expression correspond to a summation of like terms on the assumption that  $E(x_u^k)$ ,  $E(x_u^k \cdot x_v^m)$ , etc., have the same value regardless of choice-order, e.g.  $E(x_2^4) = E(x_5^4)$  or  $E(x_2^3 \cdot x_4^1) = E(x_5^3 \cdot x_2^1)$ . We have to establish the truth of this conclusion before we can employ (ii)–(v) to evaluate the moments of the non-replacement distribution.

In 12.02 we have in fact already shown that

$$E(x_u) = \mu_1; E(x_u^2) = \mu_2; E(x_u \cdot x_v) = \frac{n^2}{n^{(2)}}\mu_1 - \frac{n}{n^{(2)}}\mu_2.$$

We shall now generalise the argument employed in 12.02 to establish the conclusion that

- (a) order of choice does not affect the mean value of the  $k$ th power of the unit sub-sample score when sampling occurs without replacement;
- (b) order of choice does not affect the mean value of a co-moment of given order on the same assumption.

First, let us be clear that  $x_u^k$  is the  $k$ th power of the unit sample score at the  $u$ th draw, whereas  $x_u^k \cdot x_v^m$  is the product of the  $k$ th and  $m$ th powers of the component score of the 2-fold sub-sample consisting of an item extracted at the  $k$ th and an item extracted at the  $m$ th draw. In the same way  $x_u^k \cdot x_v^m \cdot x_w^p$  is a score referable to a 3-fold sub-sample. Thus  $x_u^3$  may be numerically equal to  $x_u^2 \cdot x_v$ , if  $x_u = x_v$  or to  $x_u \cdot x_v \cdot x_w$ , if  $x_u = x_v = x_w$ , but  $E(x_u^3)$  will not in general be numerically equal to  $E(x_u^2 \cdot x_v^1)$ .

As in 12.02 our model for what follows may be a pack containing no picture cards. Having extracted from a pack of  $n$  cards an  $r$ -fold sample which we distinguish from those which remain by the choice-order subscripts 1 to  $r$ , we may turn the remaining  $(n-r)$  cards of our model pack upwards in a row and label the score of each by its order in the sequence regardless of its numerical value, starting with  $x_{r+1}$  and ending with  $x_n$ . If we then denote the sum of the scores of this *residual universe* by  $S_{(n-r)}$ :

$$S_{(n-r)} = (x_{r+1} + x_{r+2} + \dots + x_n) = \sum_{u=r+1}^{u=n} x_u \quad \dots \quad \text{(vi)}$$

On this understanding we can identify each item in the universe of  $n$  cards; and we may denote the sum of all the score values as

$$S_n = (x_1 + x_2 + \dots + x_r + \dots + x_n) = \sum_{u=1}^{u=n} x_u \quad \dots \quad \text{(vii)}$$

In the same way we may denote as the sum of the  $k$ th power of the scores in the  $r$ -fold sample, the  $(n-r)$ -fold residual universe and the universe as a whole by

$$S_{k(r)} = \sum_{u=1}^{u=r} x_u^k; S_{k(n-r)} = \sum_{u=(r+1)}^{u=n} x_u^k; S_{kn} = \sum_{u=1}^{u=n} x_u^k \quad \dots \quad \text{(viii)}$$

In this symbolism, the  $k$ th zero moment of the *unit sample* distribution is by definition

$$\mu_k = \frac{1}{n} S_{kn} = E(x_1^k) \quad \dots \quad \text{(ix)}$$

We shall now suppose that we have already extracted without replacement an  $(a-1)$ -fold sample, so that the sum of the  $k$ th power of the scores in the residual pack of  $(n-a+1)$  cards is in this notation  $S_{k(n-a+1)}$ . The next draw is the  $a$ th and the score drawn is  $x_a$ . If we have

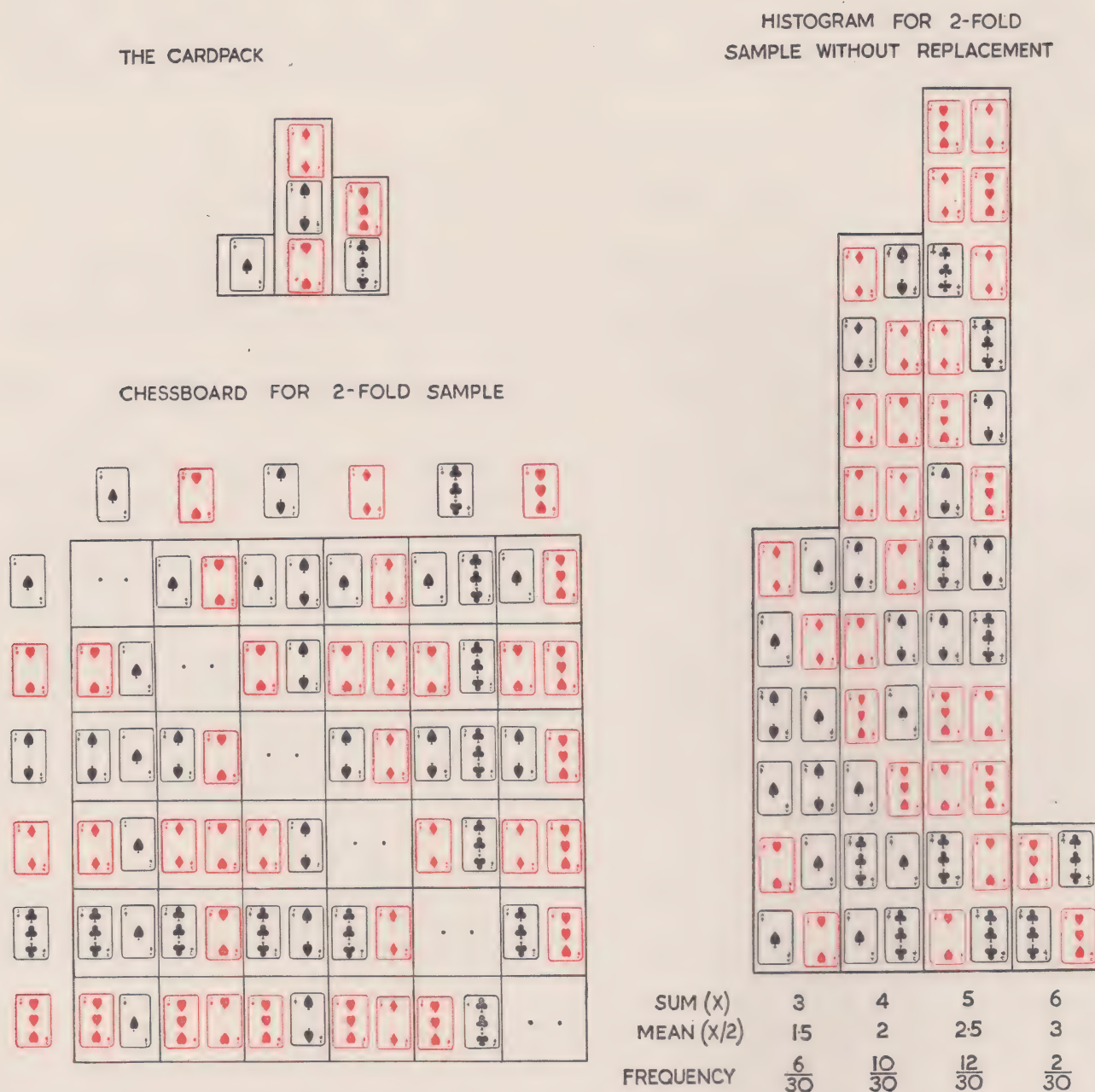


FIG. 124. Sampling without Replacement in a Finite Universe.

The universe consists of 6 items to which we attach the values 1, 2, 3 definitive of the *score class* with relative frequencies 1 : 3 : 2. In drawing an  $r$ -fold sample, we can take any item once only, but we may draw more than one item of the same score class if the class itself contains more than one. To keep track of what we have taken, we distinguish members of the same score class by *suit*. The reader may complete as an exercise the chessboard and histogram of the 3-fold sample and thus show that the half-universe sample ( $F = \frac{1}{2}$ ) has a *symmetrical* distribution. The score sums ( $X$ ) are 5, 6, 7, 8 with relative frequencies 18 : 42 : 42 : 18. For  $X = 5$  and  $X = 8$  there are 18 ways of taking the combinations 1, 2, 2 and 2, 3, 3 respectively. For  $X = 7$  and  $X = 8$  there are 36 ways of taking the combinations 1, 2, 3 and 2, 2, 3 respectively, and 6 ways of taking the combinations 2, 2, 2 and 1, 3, 3 respectively.



not replaced the sample of  $(a - 1)$  cards previously drawn, the mean value of  $x_a^k$  is the mean value of the  $k$ th power of the score from a residual universe of  $(n - a + 1)$  items, i.e.

$$E(x_a^k) = E_a(x_a^k) = \frac{E_a[S_{k(n-a+1)}]}{n - a + 1} \quad (x)$$

If we draw again, the score of the next card taken is  $x_b = x_{a+1}$ , and the sum of the  $k$ th powers of the scores in the residual universe from which we take it is  $S_{k(n-a+1)} - x_a^k$ , and this residual universe consists of  $(n - a)$  cards. For a fixed value of  $x_a$ , we may therefore write the mean value of  $x_b^k$  as

$$E_{b \cdot a}(x_b^k) = \frac{S_{k(n-a+1)} - x_a^k}{n - a}.$$

We may now employ the customary grid operation :

$$\begin{aligned} E(x_b^k) &= E_a \cdot E_{b \cdot a}(x_b^k) \\ &= \frac{E_a[S_{k(n-a+1)}]}{n - a} - \frac{E_a(x_a^k)}{n - a}. \end{aligned}$$

Whence from (x) :

$$\begin{aligned} E(x_b^k) &= \frac{E_a[S_{k(n-a+1)}]}{n - a} - \frac{E_a[S_{k(n-a+1)}]}{(n - a)(n - a + 1)}, \\ \therefore E(x_b^k) &= \frac{E_a[S_{k(n-a+1)}]}{n - a + 1} = E(x_a^k), \\ \therefore E(x_{a+1}^k) &= E(x_a^k), \\ \therefore E(x_u^k) &= E(x_1^k) = \mu_k \quad (xi) \end{aligned}$$

We have thus shown that  $E(x_u^k)$  does not depend on order of choice. For the first term in (ii)-(v) we may therefore write  $r(\mu_k)$ . To evaluate an expression of the form  $E(x_a^k \cdot x_b^m)$  we need therefore place no restriction on the value of  $b$  other than that it lies in the range 1 to  $r$  excluding  $a$  in the same range. On that understanding we write

$$E(x_a^k \cdot x_b^m) = E_a[x_a^k \cdot E_{b \cdot a}(x_b^m)].$$

In this expression  $E_{b \cdot a}(x_b^m)$  is the mean value of  $x_b^m$  associated with a fixed value of  $x_a$ , i.e. the mean value of the unit score of a pack from which we have thrown out the card whose score value is  $x_a$ . The pack so defined contains  $(n - 1)$  cards and the sum of the  $m$ th powers of the unit scores therein is  $(S_{mn} - x_a^m)$ , so that

$$\begin{aligned} E_{b \cdot a}(x_b^m) &= \frac{S_{mn} - x_a^m}{n - 1} = \frac{n}{n - 1} \mu_m - \frac{1}{n - 1} x_a^m, \\ \therefore x_a^k \cdot E_{b \cdot a}(x_b^m) &= \frac{n}{n - 1} \mu_m \cdot x_a^k - \frac{1}{n - 1} x_a^{k+m}, \\ \therefore E_a[x_a^k \cdot E_{b \cdot a}(x_b^m)] &= \frac{n}{n - 1} \mu_m \cdot E_a(x_a^k) - \frac{1}{n - 1} E_a(x_a^{k+m}), \\ \therefore E(x_a^k \cdot x_b^m) &= \frac{n}{n - 1} \mu_k \cdot \mu_m - \frac{1}{n - 1} \mu_{k+m} \quad (xii) \end{aligned}$$







Similarly, the third zero moment of the score-sum distribution is given by

$$\begin{aligned}\mu_3(S_{(r)}) &= r\mu_3 + \frac{3r^{(2)}}{n-1}[n\mu_1\mu_2 - \mu_3] + \frac{r^{(3)}}{(n-1)^{(2)}}[n^2\mu_1^3 - 3n \cdot \mu_1\mu_2 + 2\mu_3] \\ &= r\mu_3 \left[ 1 - \frac{3(r-1)}{n-1} + \frac{2(r-1)(r-2)}{(n-1)(n-2)} \right] + \frac{3nr^{(2)}\mu_1\mu_2}{(n-1)} \left[ 1 - \frac{(r-2)}{(n-2)} \right] + n^2 \frac{r^{(3)}}{(n-1)^{(2)}}\mu_1^3 \\ &= \frac{r(n-r)(n-2r)}{(n-1)^{(2)}}\mu_3 + \frac{3n(n-r)r^{(2)}}{(n-1)^{(2)}}\mu_1\mu_2 + \frac{n^2r^{(3)}}{(n-1)^{(2)}}\mu_1^3.\end{aligned}$$

Whence we derive by the customary conversion formula

$$\begin{aligned}m_3(S_{(r)}) &= \frac{r(n-r)(n-2r)}{(n-1)^{(2)}}[\mu_3 - 3\mu_1\mu_2 + 2\mu_1^3], \\ \therefore m_3(S_{(r)}) &= \frac{r(n-r)(n-2r)}{(n-1)^{(2)}} \cdot m_3 \cdot \cdot \cdot \cdot \cdot \quad (\text{xvii})\end{aligned}$$

In the same way, we get the fourth zero moment of the  $r$ -fold sampling distribution of the score-sum

$$\begin{aligned}\mu_4(S_{(r)}) &= r\mu_4 + \frac{4r^{(2)}}{n-1}[n\mu_1\mu_3 - \mu_4] + \frac{3r^{(2)}}{n-1}[n\mu_2^2 - \mu_4] \\ &\quad + \frac{6r^{(3)}}{(n-1)^{(2)}}[n^2\mu_1^2\mu_2 - 2n\mu_1\mu_3 - n\mu_2^2 + 2\mu_4] \\ &\quad + \frac{r^{(4)}}{(n-1)^{(3)}}[n^3\mu_1^4 - 6n^2\mu_2\mu_1^2 + 3n\mu_2^2 + 8n\mu_1\mu_3 - 6\mu_4] \\ &= \frac{r(n-r)}{(n-1)^{(3)}}[(n-2r)(n-3r) - n(r-1)]\mu_4 \\ &\quad + \frac{4nr^{(2)}}{(n-1)^{(3)}}[(n-2)(n-3) - 3(r-2)(n-3) + 2(r-2)(r-3)]\mu_1\mu_2 \\ &\quad + \frac{6n^2r^{(3)}(n-r)}{(n-1)^{(3)}}\mu_1^2\mu_2 + \frac{n^3r^{(4)}}{(n-1)^{(3)}}\mu_1^4 \\ &\quad + \frac{3nr^{(2)}}{(n-1)^{(3)}}[(n-2)(n-3) - 2(r-2)(n-3) + (r-2)(r-3)]\mu_2^2 \\ &= \frac{r(n-r)}{(n-1)^{(3)}}[(n-2r)(n-3r) - n(r-1)]\mu_4 \\ &\quad + \frac{4n(n-r)(n-2r+1)r^{(2)}}{(n-1)^{(3)}}\mu_1\mu_3 + \frac{6n^2r^{(3)}(n-r)}{(n-1)^{(3)}}\mu_1^2\mu_2 \\ &\quad + \frac{n^3r^{(4)}}{(n-1)^{(3)}}\mu_1^4 + \frac{3n \cdot r^{(2)}(n-r)(n-r-1)}{(n-1)^{(3)}}\mu_2^2.\end{aligned}$$





for all values of  $F$  if  $\beta_1 = 0$ ; and  $\beta_2(S_{(r)})$  may be greater or less than 3. If  $F$  is small and  $n$  is large, (xxi)-(xxii) approach the limit for the replacement set-up, as we should expect, i.e.

$$\beta_1(S_{(r)}) \simeq \frac{1}{r}\beta_1 \quad \text{and} \quad \beta_2(S_{(r)}) \simeq 3 + \frac{\beta_2 - 3}{r}.$$

### 24-fold Symmetrical Universes of 3 classes.

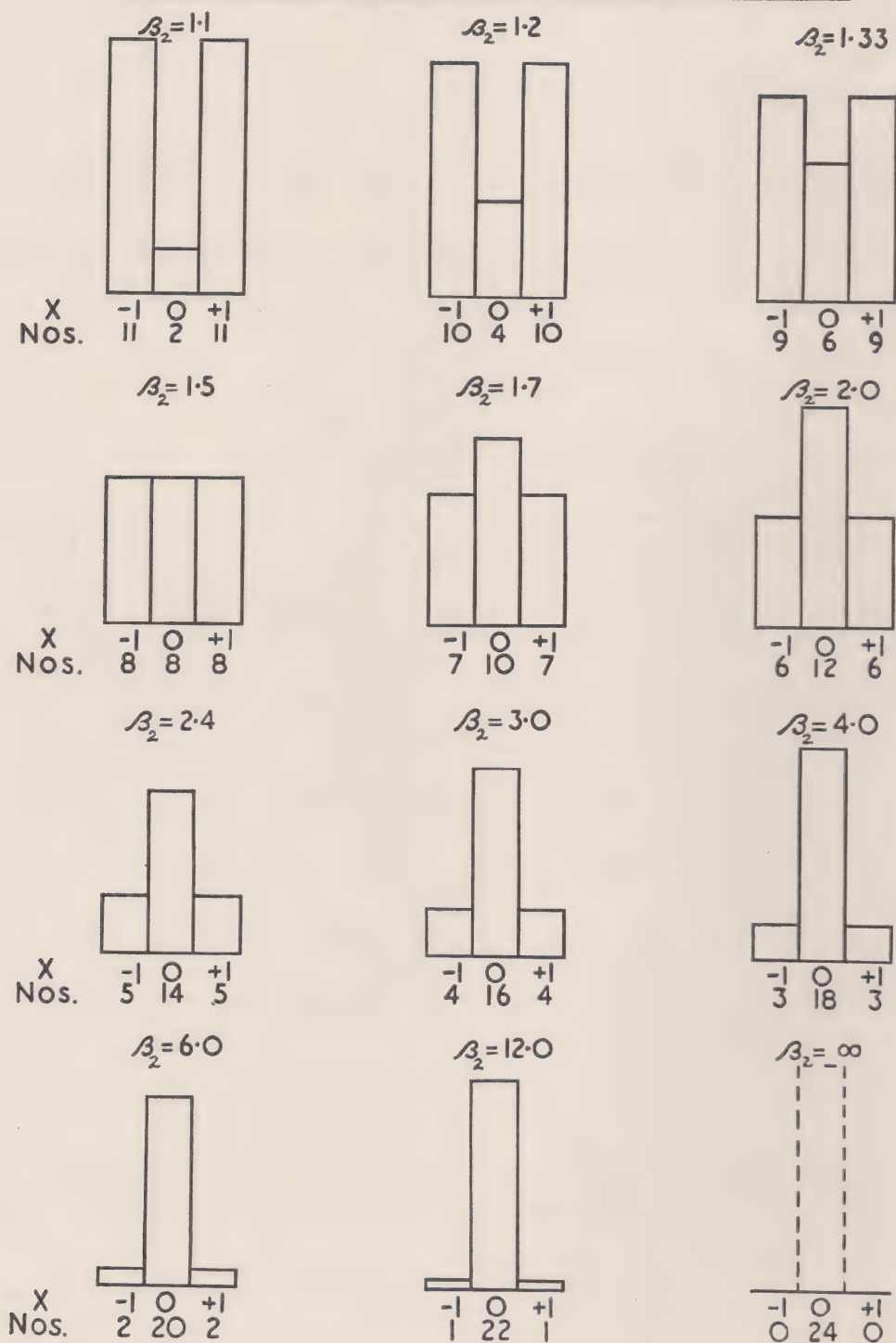


FIG. 125. Symmetrical 24-fold Universes of 3 classes with kurtosis ( $\beta_2$ ) coefficients.

Note the scale is not uniform throughout.



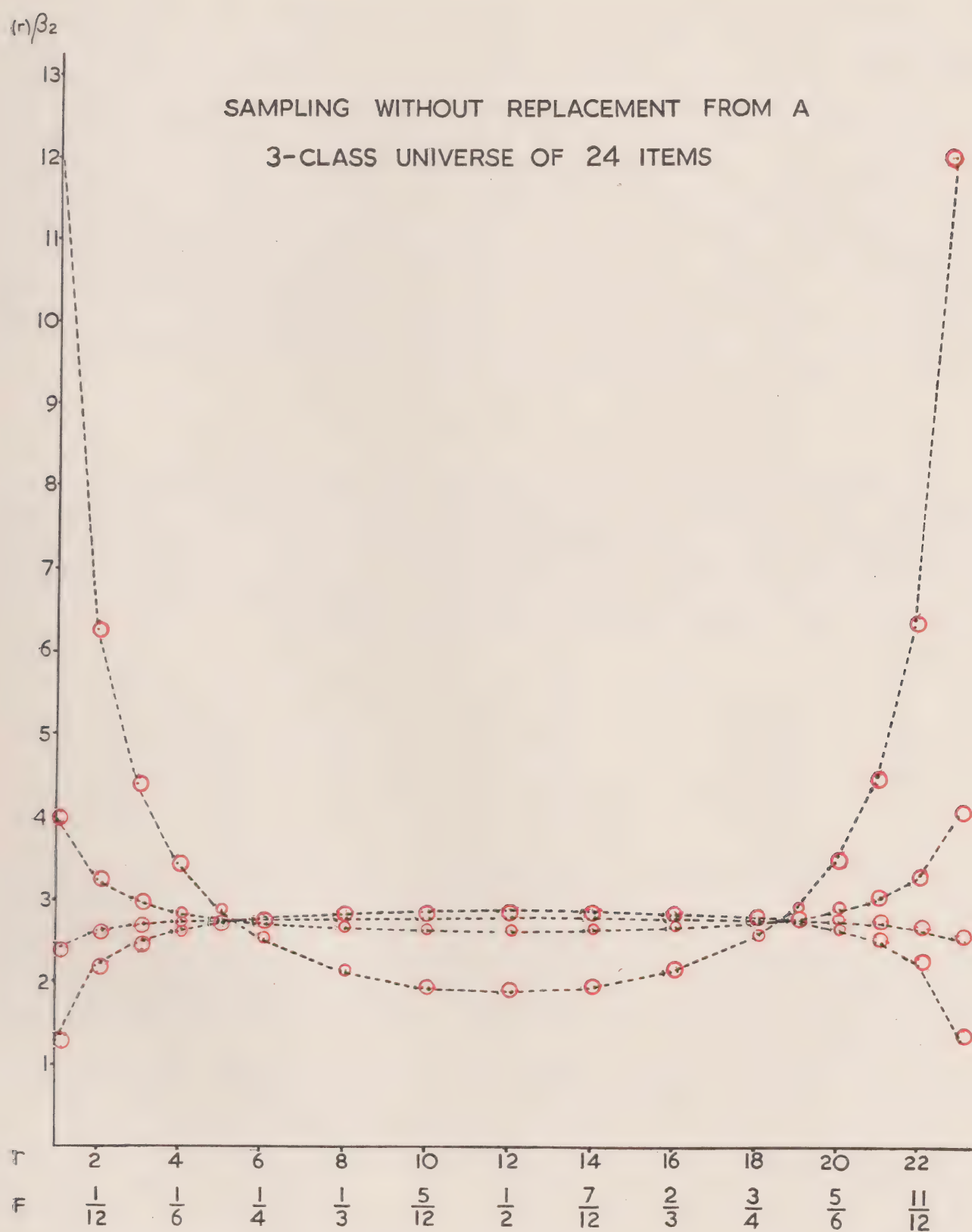


FIG. 126. Variation of kurtosis coefficient ( $\beta_2$ ) with size of sampling fraction for 24-fold universes of 3 classes shown in Fig. 125.

For small values of  $F$ , we therefore expect that the score-sum distribution will be leptokurtic only if that of the unit sample is also ; but if  $F$  is large, the score-sum distribution may be leptokurtic if that of the unit sample is platykurtic. In the limit, of course,  $\beta_2(S_{(r)})$  becomes infinite, since only one value ( $S_n$ ) of the score-sum is consistent with exhaustive sampling.

A closer examination of the approximate formula (xxii) brings into focus what is perhaps a more remarkable feature of non-replacement sampling distribution than the symmetry of the half-universe sample. The expression  $1 - 6F(1 - F)$  vanishes when  $F = \frac{1}{2} \pm 1/\sqrt{12}$ , i.e.  $F \simeq 0.22$  or  $0.79$  between which limits the coefficient of  $\beta_2$  in the second term of (xxii) is negative with a numerical maximum for  $F \simeq 0.59$ . Regardless of the structure of the universe, the sampling distribution will always be platykurtic unless the sampling fraction in round numbers is greater than four-fifths ; *ceteris paribus*, below this level a highly leptokurtic unit-sample distribution will generate a more platykurtic sampling distribution than a distribution which is initially flatter than the normal curve, e.g. a rectangular or indeed even a  $U$ -shaped one. At  $F \simeq 0.59$  the kurtosis is a minimum and (xxii) is approximately

$$3 - \frac{3 + 1.1\beta_2}{r}.$$

This seemingly paradoxical characteristic of any non-replacement distribution comes into focus (Fig. 125) if we calculate from the exact formula (xx) the kurtosis for samples of different sizes extracted without replacement from a symmetrical 24-fold universe of only 3 score classes. For simplicity, we may assign to the 3 classes scores of  $-1$ ,  $0$  and  $+1$ , and frequencies ( $p_a, p_b, p_c$ ) as below with  $\beta_2$  values for the unit-sample distribution in the range 1–12 including a rectangular and a  $U$ -shaped contour at the lower limit of kurtosis.

$-1$	$0$	$+1$	$\beta_2$	$-1$	$0$	$+1$	$\beta_2$
1	21	1	12.0	5	14	5	2.5
2	20	2	6.0	6	12	6	2.0
3	18	3	4.0	8	8	8	1.3
4	16	4	3.0	11	2	11	1.1

The picture disclosed by Fig. 125 raises the question : what lower limit may  $\beta_2(S_{(r)})$  attain if  $\beta_2$  is as high as may be ? This admits no simple answer because the size of the universe itself sets a limit both to the maximum value of  $\beta_2$  and to the value of  $r$  consistent with the condition that the second term in (xxii) is both negative and numerically maximal, as when  $F \simeq 0.6$ . Thus a universe of 100 items assignable to 3 equally spaced score classes ( $-1, 0, +1$ ) as above cannot have a kurtosis greater than 50, and the sample size consistent with a minimum value of  $\beta_2(S_{(r)})$  is about 60.

For the rectangular universe  $\beta_1 = 0$  and  $\beta_2 \simeq 1.8$  when  $n$  is large. The possibility of generating a rectangular sampling distribution therefore implies that  $3 + 1.1\beta_2 = 1.2r$ . If  $n = 200$  and  $F = 0.6$ ,  $r = 120$ , so that  $\beta_2$  could satisfy this relation only if  $\beta_2 \simeq 130$ . For the 200-fold binomial universe defined by  $(0.995 + 0.005)^1$ ,  $\beta_2$  exceeds 130 but no other binomial 200-fold universe and no 200-fold universe of more than 2 non-zero classes can satisfy the condition  $\beta_2 \geq 130$ . From a 2-class universe of 1 zero score value and 199 unit score values the value of  $\beta_2(S_{(r)})$  for the 120-fold sample would be 1.12 ; but the sample itself would contain only 2 score classes (*viz.* score sums of 120 and 119) as we see by expanding  $(199 + 1)^{120}$ . Though  $\beta_2(S_{(r)})$  is in this case less than 1.8, the distribution of the sample score is monotonic.



From (xix)–(xx) we see that the first two Pearson coefficients of the  $r$ -fold and the  $(n - r)$ -fold sample are respectively identical. Thus  $\beta_1(S_{(r)}) = \beta_1$  and  $\beta_2(S_{(r)}) = \beta_2$  when  $r = (n - 1)$ . If  $\beta_2$  lies in the neighbourhood of  $3(n - 1) \div (n + 1)$ , the kurtosis of the  $r$ -fold distribution does not appreciably change within the range  $r = 1$  to  $r = n - 1$ , e.g. when  $n = 24$  and  $\beta_2 = 2.76$  (Fig. 126).

### 19.05 DIFFERENCE DISTRIBUTION FOR NON-REPLACEMENT SAMPLING

To derive the distribution of the difference between the raw scores of an  $a$ -fold and  $b$ -fold sample from the same finite universe we must retrace our steps to the derivation of (ix)–(xii) in 19.03. We there considered the form of the terms of the expansion of a multinomial expression such as  $(p + q + r + s \dots)^n$  containing  $m$  basic terms  $p, q, r$ , etc. all positive. For the  $(a + b)$ -fold score-sum the expressions  $K_{2,0}$ ,  $K_{1,1}$ , etc. are obtainable by inserting  $(a + b)$  for  $m$ . We may write the raw-score difference in the form

$$\begin{aligned} D &= (x_{1,a} + x_{2,a} \dots + x_{a,a}) - (x_{1,b} + x_{2,b} \dots + x_{b,b}) \\ &= (x_{1,a} + x_{2,a} \dots + x_{a,a} - x_{1,b} - x_{2,b} \dots - x_{b,b}), \end{aligned}$$

whence the  $k$ th zero moment is

$$(a - b)\mu_k = E(x_{1,a} + x_{2,a} \dots - x_{1,b} - x_{2,b} \dots)^k.$$

The expression on the right has  $a$  positive and  $b$  negative terms within the brackets; and we may classify the terms of the expansion as in the derivation of (ix)–(xii) of 19.03. The coefficients of corresponding classes will not be identical with  $K_{2,0}$ , etc. Accordingly, we shall label them as  $H_{2,0}$ , etc. The reader should first note that the correct interpretation of  $(a - b)^{(r)}$  is consistent with Vandemonde's formula, if we write it as

$$(a - b)^{(r)} = \sum_{k=1}^{k=r} r_{(k)} a^{(k)} (-b)^{(r-k)}.$$

This would be strictly analogous to the ordinary binomial expansion  $(a - b)^r$  if it were true that  $(-b)^{(k)} = (-1)^k \cdot b^{(k)}$ . For brevity, it is convenient to define by use of square brackets an expression which precisely conforms to the analogy, viz.:

$$(a - b)^{[r]} = \sum_{k=1}^{k=r} (-1)^k r_{(k)} \cdot a^{(k)} \cdot b^{(r-k)}.$$

The reader may check that the pattern for the  $H$  coefficients is as follows:

$$\begin{array}{ll} H_{2,0} = (a + b), & H_{4,0} = (a + b), \\ H_{1,1} = (a - b)^{[2]}, & H_{3,1} = 4(a - b)^{[2]}, \\ H_{3,0} = (a - b), & H_{2,2} = 3(a + b)^{(2)}, \\ H_{2,1} = 3a^{(2)} - 3b^{(2)}, & H_{2,1,1} = 6[a^{(3)} - ba^{(2)} - ab^{(2)} + b^{(3)}], \\ H_{1,1,1} = (a - b)^{[3]}, & H_{1,1,1,1} = (a - b)^{[4]}. \end{array}$$

We may then derive, by recourse to the  $H$  coefficients defined above, compact expressions for the moments of the raw-score difference distribution in terms of the moments of the u.s.d.:

$$\begin{aligned} (a - b)\mu_1 &= (a - b)\mu_1, \\ (a - b)\mu_2 &= (a + b)\mu_2 + (a - b)^{[2]}\mu_1^2, \\ (a - b)\mu_3 &= (a - b)\mu_3 + 3(a^{(2)} - b^{(2)})\mu_2 \cdot \mu_1 + (a - b)^{[3]}\mu_1^3, \\ (a - b)\mu_4 &= (a + b)\mu_4 + 4(a - b)^{[2]}\mu_3 \cdot \mu_1 + 3(a + b)^{(2)}\mu_2^2 + \\ &\quad 6(a^{(3)} - ba^{(2)} - ab^{(2)} + b^{(3)})\mu_2 \cdot \mu_1^2 + (a - b)^{[4]}\mu_1^4. \end{aligned}$$

We thus derive the following expressions for the first two Pearson coefficients of the raw-score difference distribution in terms of those of a u.s.d. referable to  $n$  items:

$${}_{(a-b)}\beta_1 = \frac{(a-b)^2 \{n^2 - 3n(a+b) + 2(a-b)^2\}^2 (n-1)}{\{n(a+b) - (a-b)^2\}^3 (n-2)^2} \beta_1$$

$${}_{(a-b)}\beta_2 = \frac{(n-1)[s(n-s)\{n(n+1) - 6s(n-s)\} + 16ab\{n(n+1) - 3s(n-s) - 6ab\}]}{(n-2)^{(2)}\{s(n-s) + 4ab\}^2} \beta_2$$

$$+ \frac{3n^{(2)}[s^{(2)}(n-s)^{(2)} + 8ab\{s(n-s) + 2(ab-n+1)\}]}{(n-2)^{(2)}\{s(n-s) + 4ab\}^2}$$

in which  $s = a + b$ .

Both expressions simplify greatly, if we choose samples of equal size ( $a = b$ ), in which event

$${}_{(a-b)}\beta_1 = 0$$

and

$${}_{(a-b)}\beta_2 = \frac{(n-1)}{2an(n-2)^{(2)}} \{n(n-6a) + n + 6a\} \beta_2$$

$$+ \frac{3(n-1)}{2an(n-2)^{(2)}} \{(2a-1)(n-2a)^{(2)} + 8a^2(n-a) - 8a(n-1)\}$$

If  $a = n(n+1) \div 6(n-1) = b$ , it is thus apparent that the difference distribution is symmetrical; and the value of the second Pearson coefficient is independent of  $\beta_2$ , i.e. of the structure of the universe. On substitution of this sample size in the expression above we find that  ${}_{(a-b)}\beta_2$  reduces to  $3(n-1) \div (n+1)$ ; but the interpretation of this result is meaningful only within the framework of the assumption that both  $n$  and  $a$  must be integers. Evidently the coefficient of  $\beta_2$  will be small if  $n = 6a$ , i.e. *each* sample is a one-sixth fraction of the universe of choice. For large values of  $n$  we may thus say that an *overall* sampling fraction of one-third will ensure that the kurtosis of the difference distribution is independent of the kurtosis of the u.s.d. More generally for  $n = 6a$ ,  ${}_{(a-b)}\beta_2$  reduces to

$$\frac{6(n-1)^2}{n^{(4)}} \beta_2 + \frac{3(n-1)^2(n^2 - 6n + 6)}{n^{(4)}}.$$

The maximum finite value of  $\beta_2$  occurs in the binary universe, the frequencies of the classes being  $n^{-1}$  and  $(n-1)n^{-1}$  respectively, one class being then represented by only one number. The second Pearson coefficient of its u.s.d. is  $(n^2 - 3n + 3) \div (n-1)$ . This is its maximum value; and the maximum value of  ${}_{(a-b)}\beta_2$  is therefore exactly 3. Thus the difference distribution is necessarily platykurtic and the greatest contribution which can be made by the term involving  $\beta_2$  is  $6(n^2 - 3n + 3) \div n(n-2)^2$ . The table below shows for various values of  $n$  the values of the two terms in  ${}_{(a-b)}\beta_2$  assumption that  $\beta_2$  has its maximum value, as above.

$n$	1st term	2nd term	$n$	1st term	2nd term
6	1.75	1.25	30	0.21	2.79
12	0.62	2.38	42	0.15	2.35
18	0.38	2.62	60	0.10	2.90
24	0.27	2.73	96	0.06	2.94



Even if the u.s.d. of the binary universe is very platykurtic, we therefore see that samples of size equal to  $\frac{1}{6}$ th of the universe will generate a symmetrical difference distribution having a second Pearson coefficient greater than or equal to 2.8 if  $n$  is greater than or equal to 30. For any universe of 30 or more items, regardless of the number of score classes and of items in each, there is good reason to assume that the first two Pearson coefficients of the distribution of the difference referable to equal samples of  $\frac{1}{6}$ th will lie close to their normal values. However, this does not suffice to justify the conclusion that the normal curve will give an adequate quadrature for the sample difference distribution. An examination of how gratuitous such an assumption may be will indeed give us some insight into circumstances which guarantee a good fit.

In particular, we recall the case of sampling from a 2-class universe. Without restriction on the values of  $a$  and  $b$ , the difference distribution is then definable as follows for a u.s.d. of score values differing by unit increment

Difference Scores	.	.	.	- 1	0	+ 1
Frequencies	.	.	.	$\frac{a}{n}$	$\frac{n - a - b}{n}$	$\frac{b}{n}$

When  $a = b$  and  $F = (a + b) \div n$  is the total sampling fraction, this reduces to

Difference Scores	.	.	.	- 1	0	+ 1
Frequencies	.	.	.	$\frac{F}{2}$	$1 - F$	$\frac{F}{2}$

Whence  ${}_{(a-b)}\beta_1 = 0$  and  ${}_{(a-b)}\beta_2 = 3.0$  if  $a = b$  and  $F = \frac{1}{3}$ .

The difference distribution is then a special case of what we have elsewhere called (14.05) the *burette universe*, viz.:

Score	.	.	.	- 1	0	+ 1
Frequency	.	.	.	$\frac{1}{6}$	$\frac{2}{3}$	$\frac{1}{6}$

We may here make use of results obtainable (14.07) from sampling in the burette (infinite discrete 3-class) universe by stating at this stage without proof the following conclusion: when the first two Pearson coefficients of a distribution involving 20 score classes are very close to their normal values, we may confidently invoke the normal distribution for purposes of quadrature adequate for statistical usage. It is therefore immaterial to examine the implications of the foregoing formulae for  ${}_{(a-b)}\beta_1$  and  ${}_{(a-b)}\beta_2$  more closely. It suffices to state of any finite unimodal universe that:

(a) the first 2 coefficients of the non-replacement difference distribution w.r.t.  $a$ -fold and  $b$ -fold samples will lie very close to their normal values if both the following conditions hold good:

- (i) the sample sizes are equal ( $a = b$ );
- (ii) the total sampling fraction ( $F = (a + b) \div n$ ) is in the neighbourhood of one-third;

(b) the normal curve will then give a satisfactory quadrature if the distribution of the  $a$ -fold sample is referable to at least 10 different score values.

## 19.06 THE SO-CALLED CHI-SQUARE STATISTIC

In Vol. I we have sufficiently clarified the distinction between two methods of scoring respectively referred to throughout this book as *taxonomic* and *representative*. When we score by the former method we specify a sample by the number of individuals in each of an exhaustive set of exclusive classes. When the classification is binary every individual belongs to class *A* or to class *B*. If there are *a* individuals of an *r*-fold sample belonging to class *A*, there must therefore be  $(r - a) = b$  individuals in class *B*. Conversely,  $a = (r - b)$ ; and one score suffices to define a sample of known size. For example, it is immaterial whether we specify a 12-fold sample of peas classified as green and yellow by the fact that it contains 5 green or 7 yellow.

When our concern is with more than two classes, this is not so. If there are *N* classes, our knowledge of the *r*-fold is not complete unless we can specify the score of  $(N - 1)$  classes. For instance, we can exhaustively specify one and the same flock of 25 Andalusian fowls classified as white, black and blue in three different ways, *viz.* :

(i)	(ii)	(iii)
White 12	White 12	Black 7
Black 7	Blue 6	Blue 6

To avoid periphrasis, we may speak of a classification involving more than 2 classes as *manifold*. The problem we shall examine in this chapter is the correspondence between hypothesis and expectation in a manifold system. Thus we might wish to know whether the composition of the 25-fold sample cited above is statistically consistent with the requirements of the Mendelian ratio : 1 : 2 : 1 for white, blue and black respectively.

We can, of course, specify the probability (*P*) of getting a sample of a given composition by recourse to the multinomial theorem in ordinary (*replacement*) or factorial (*non-replacement*) powers. For our Andalusian flock the data are

	White	Blue	Black	Total
Observed numbers . . .	12	6	7	25
Unit-sample expectation . .	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	1

On the assumption that the universe is indefinitely large, we consider that we are sampling with replacement and put

$$P = \frac{25!}{12!6!7!} \left(\frac{1}{4}\right)^{12} \left(\frac{1}{2}\right)^6 \left(\frac{1}{4}\right)^7.$$

In dealing with a 2-class system, we commonly specify unit sample expectation w.r.t. choice of an item of class *B* as  $q = (1 - p)$ , that of a choice of a single item of class *A* being *p*. In a manifold system of more than 2 classes, no such unique relation exists between the unit sample expectation w.r.t. class *A* and to class *B*. Accordingly, we shall write the unit sample distribution for a system of *N* classes as

$$p_a + p_b + p_c \dots + p_n.$$

In this expression  $p_a$  is the unit sample expectation w.r.t. class *A*, and  $(1 - p_a) = q_a$  is the expectation that a choice of a single item will *not* belong to class *A*. Where occasion arises, we may write  $q_b = (1 - p_b)$ ,  $q_c = (1 - p_c)$ , etc. For a 3-class system therefore

$$p_c = (1 - p_a - p_b) \quad \text{and} \quad q_c = p_a + p_b \quad . \quad . \quad . \quad . \quad (i)$$









a unit sample will *not* belong to class  $S$ , if  $p_s$  is the probability that it will do so ; and  $rp_s(1 - p_s)$  is the expected value of the square score deviation  $(x_s - M_s)^2$ , i.e. the variance of the distribution of the  $S$ -score. Hence we may write

$$E(x_s - M_s)^2 = rp_s(1 - p_s).$$

Since  $M_s = rp_s$

$$E(C_N) = \sum_{s=1}^{s=N} (1 - p_s) = N - \sum_{s=1}^{s=N} p_s.$$

By definition

$$(p_a + p_b + p_c \dots p_N) = 1,$$

$$\therefore E(C_N) = N - 1. \quad \dots \dots \dots (xi)$$

Thus the statistic  $C_2$  of the binary system is a particular case of a more general pattern which takes account of score deviations of *all* the constituent  $N$  classes, and the expected value of this statistic is that of a *Chi-Square* variate of  $N - 1$  degrees of freedom. For the 3-class replacement system whose general term is (x) above

$$C_3 = \frac{u^2}{M_a} + \frac{v^2}{M_b} + \frac{w^2}{M_c} \quad \dots \dots \dots (xii)$$

In this case,  $N = 3$  and  $(N - 1) = 2 = E(C_3)$ ; and we shall later explore the possibility that  $C_3$  is in fact *approximately* expressible as a *Chi-Square* variate of 2 d.f. The procedure will make it sufficiently clear that the rule suggested holds good when  $N > 3$ .

It will clarify our task if we first re-examine the implications of the statement that the square standard score deviation ( $C_2$ ) of a binomial distribution has approximately the distribution of *Chi-Square* for 1 d.f. We have previously arrived at this conclusion by the following route :

- (i) The distribution of the score deviation  $(a - rp_a) = u$  tallies closely with that of a normal variate with variance  $rp_a(1 - p_a)$  for large values of  $r$  unless  $p_a$  or  $1 - p_a$  is small compared with the reciprocal of  $r$ , i.e. for large values of  $r$  and  $rp_a > 10$  ;
- (ii) Subject to the qualifications stated, the ratio ( $c$ ) of  $u$  to  $\sqrt{rp_a(1 - p_a)}$  is therefore approximately a normal variate of unit variance ;
- (iii) Since the square of a normal variate of unit variance is a *Chi-Square* variate of 1 d.f., the ratio of  $u^2$  to  $rp_a(1 - p_a)$  is also approximately within the framework of the same qualifications a *Chi-Square* variate of 1 d.f.

Let us be clear that we are *not* speaking of the *exact* distribution of  $C_2 = c^2$  in these terms. Accordingly, we might regard the problem as that of finding a good fitting curve for it by the method of moments. Now we have seen in 13.01-13.02 that the following relations hold good for the mean moments of the normal variate ( $c$ ) of unit variance, and the zero moments of the *Chi-Square* variate ( $C$ ) for 1 d.f. :

$$\mu_k(C) = m_{2k}(c).$$

In this expression, we have seen that  $m_2(c) = 1$ ,  $m_4(c) = 3$ ,  $m_6(c) = 15$  and  $m_8(c) = 105$ . Hence

$$\mu_1(C) = 1 ; \mu_2(C) = 3 ; \mu_3(C) = 15 ; \mu_4(C) = 105 \quad \dots \dots \dots (xiii)$$

If  $C_2$  defined by (ix) above is a statistic for which we seek a fitting curve, we may proceed to determine its moments as follows :

$$\mu_2(C_2) = E(C_2^2) = E\left(\frac{u^2}{M_a} + \frac{v^2}{M_b}\right)^2 = \frac{E(u^4)}{M_a^2} + \frac{E(v^4)}{M_b^2} + \frac{2E(u^2 \cdot v^2)}{M_a M_b}.$$

Since  $u^2 = v^2$  in a 2-class system, we may here put

$$\mu_2(C_2) = \frac{E(u^4)}{M_a^2 M_b^2} (M_a^2 + M_b^2 + 2M_a M_b) = \frac{E(u^4)(M_a + M_b)^2}{M_a^2 M_b^2}.$$

In this expression  $E(u^4)$  is the mean value of the 4th power of the raw-score distribution of the binomial distribution, i.e. its 4th mean moment ( $m_4$ ). Also

$$(M_a + M_b)^2 = r^2 \quad \text{and} \quad M_a^2 M_b^2 = r^4 p_a^2 (1 - p_a)^2,$$

$$\therefore \frac{(M_a + M_b)^2}{M_a^2 M_b^2} = \frac{1}{r^2 p_a^2 (1 - p_a)^2} = \frac{1}{m_2^2},$$

$$\therefore \mu_2(C_2) = \frac{m_4}{m_2^2} = \beta_2.$$

In 14.05 we have obtained the value of  $\beta_2$  for the  $r$ -fold sample from a universe whose unit sample distribution is  $(q + p)^1$ , viz. :

$$3 + \frac{1 - 6pq}{rpq}.$$

Evidently, the second zero moment of the  $C_2$  distribution tends to 3 when  $r$  is large and the reciprocal of either  $p_a$  or  $p_b$  is small in comparison with  $r$ . In the same way, we may see that  $\mu_3(C_2)$  and  $\mu_4(C_2)$  approach the values of  $\mu_k(C)$  defined by (xiii) above. The numerical values 1, 3, 15, 105, for the particular case  $f = 1$ , i.e. the Chi-Square variate of 1 d.f., illustrates the more general rule :

$$\mu_1 = f; \quad \mu_2 = f(f + 2); \quad \mu_3 = f(f + 2)(f + 4);$$

$$\mu_4 = f(f + 2)(f + 4)(f + 6), \text{ etc.}$$

The statistic defined by (xii) is referable to a 3-class system, and the hypothesis we are exploring is that its approximate replacement distribution for large samples is that of a Chi-Square variate of  $(3 - 1) = 2$  d.f. If  $f = 2$  in the above

$$\mu_1 = 2; \quad \mu_2 = 8; \quad \mu_3 = 48; \quad \mu_4 = 384 \quad . \quad . \quad . \quad . \quad (xiv)$$

We have already seen that  $\mu_1 = 2$ , when  $N = 3$ . We shall now explore the possibility that the limiting values of  $\mu_2$ , etc., for large values of  $r$  conform with the above.

Before proceeding further, we may pause to refer to an ambiguity of current terminology. Pearson developed the theory of the distribution referred to as the Chi-Square distribution on the assumption of continuous variation; and it is in that sense that we speak of a Chi-Square variate elsewhere in this volume. The exact distribution of the function of sums of squares defined by (x) is necessarily *discrete*. Accordingly, it is misleading to speak of  $C_N$  in (x) as a Chi-Square variate, and *a fortiori* misleading to define a Chi-Square variate as a sum of squares so weighted. What we can say is what we have already found reasons for suspecting, viz. that the Chi-Square for  $(N - 1)$  degrees of freedom is a good fitting curve for the sampling distribution of  $C_N$  when  $r$  is very large. Only on that understanding can we use the table of the appropriate Chi-Square integral with propriety to evaluate the expectation that  $C_N$  will exceed a certain numerical value.



## 19.07 THE MOMENTS OF THE SO-CALLED CHI-SQUARE STATISTIC

We may write the statistic defined by (xii) in 19.06 in the form :

$$\begin{aligned} C_3 &= \frac{(a - M_a)^2}{M_a} + \frac{(b - M_b)^2}{M_b} + \frac{(c - M_c)^2}{M_c} \\ &= \frac{a^2}{M_a} + \frac{b^2}{M_b} + \frac{c^2}{M_c} - 2(a + b + c) + (M_a + M_b + M_c) \\ &= \frac{a^2}{M_a} + \frac{b^2}{M_b} + \frac{c^2}{M_c} - r, \\ \therefore C_3 + r &= \frac{a^2}{M_a} + \frac{b^2}{M_b} + \frac{c^2}{M_c}. \end{aligned}$$

Whence we may put

$$\begin{aligned} E(C_3 + r)^2 &= E\left(\frac{a^2}{M_a} + \frac{b^2}{M_b} + \frac{c^2}{M_c}\right)^2, \\ \therefore E(C_3^2) + 2rE(C_3) + r^2 &= \frac{E(a^4)}{M_a^2} + \frac{E(b^4)}{M_b^2} + \frac{E(c^4)}{M_c^2} \\ &\quad + \frac{2E(a^2b^2)}{M_aM_b} + \frac{2E(a^2c^2)}{M_aM_c} + \frac{2E(b^2c^2)}{M_bM_c}. \end{aligned}$$

In the preceding expression,  $E(C_3) = 2$ , being the mean value of  $C_3$  as already shown. The expression  $E(a^4)$  is the 4th zero moment of the  $A$  class score, and we may write it accordingly as  $\mu_4(a)$ . For brevity we may also write

$$E(a^2b^2) = \mu_{2.2.0}; \quad E(a^2c^2) = \mu_{2.0.2}; \quad E(b^2c^2) = \mu_{0.2.2}.$$

Thus the foregoing expression reduces to

$$\mu_2(C_3) = \frac{\mu_4(a)}{M_a^2} + \frac{\mu_4(b)}{M_b^2} + \frac{\mu_4(c)}{M_c^2} + \frac{2\mu_{2.2.0}}{M_aM_b} + \frac{2\mu_{2.0.2}}{M_aM_c} + \frac{2\mu_{0.2.2}}{M_bM_c} - r^2 - 4r \quad (i)$$

The hypothesis that the Chi-Square distribution for 2 d.f. is a satisfactory fitting curve for the 3-class statistic  $C_3$  requires *inter alia* that

$$\mu_2(C_3) \simeq f(f+2) = 8 \quad . \quad . \quad . \quad . \quad . \quad (ii)$$

To evaluate the variance of the distribution of  $C_3$  as defined by (i) it is necessary to find expressions for  $\mu_{2.2.0}$ , etc. More generally, the evaluation of higher moments presupposes that we can find expressions for

$$\mu_{h.i.j} = E(a^h \cdot b^i \cdot c^j).$$

Since we can express  $c = (r - a - b)$  in terms of  $a$ ,  $b$  and  $r$  the fixed size of the sample, we can always transform co-moments of the above form to the simpler pattern illustrated by the following :

$$\begin{aligned} \mu_{2.2.2} &= E[a^2b^2(r - a - b)^2] \\ &= E[a^2b^2(r^2 + a^2 + b^2 - 2ra - 2rb + 2ab)] \\ &= r^2E(a^2b^2) - 2rE(a^3b^2) - 2rE(a^2b^3) + 2E(a^3b^3) + E(a^4b^2) + E(a^2b^4), \\ \therefore \mu_{2.2.2} &= r^2\mu_{2.2.0} - 2r\mu_{3.2.0} - 2r\mu_{2.3.0} + 2\mu_{3.3.0} + \mu_{4.2.0} + \mu_{2.4.0} \quad (iii) \end{aligned}$$

It will thus suffice for the purpose of evaluating moments of any order, if we define  $\mu_{h.i.o}$ .

We can do this directly by recourse to the grid symbolism of 11.01–11.04 employed in our treatment of the two card-pack model of 12.03; but it will be instructive if we also perform the operation by recourse to first principles. The general term of the distribution which defines the  $r$ -fold sample frequency of the particular score values  $a$ ,  $b$  and  $c$  is

$$\frac{r!}{a! b! c!} p_a^a \cdot p_b^b \cdot p_c^c.$$

If we write  $p_a = (1 - q_a)$ ,

$$1 - \frac{p_b}{q_a} = \frac{1 - p_b - p_a}{q_a} = \frac{p_c}{q_a} \quad \text{and} \quad q_a^b \cdot q_a^c = q_a^{r-a}.$$

Hence we may put

$$\frac{r!}{a! b! c!} p_a^a \cdot p_b^b \cdot p_c^c = \frac{r!}{a! (r-a)!} p_a^a q_a^{r-a} \cdot \frac{(r-a)!}{b! c!} \left(\frac{p_b}{q_a}\right)^b \left(\frac{p_c}{q_a}\right)^c. \quad (\text{iv})$$

By definition, the mean value of  $a^h b^i$  is given by

$$E(a^h b^i) = \sum_{a=0}^{a=r} \sum_{b=0}^{b=r-a} a^h b^i \cdot \frac{r!}{a! b! c!} p_a^a \cdot p_b^b \cdot p_c^c.$$

Whence by (iv)

$$E(a^h b^i) = \sum_{a=0}^{a=r} a^h \cdot \frac{r!}{a! (r-a)!} p_a^a q_a^{r-a} \cdot \sum_{b=0}^{b=r-a} b^i \frac{(r-a)!}{b! c!} \left(\frac{p_b}{q_a}\right)^b \left(\frac{p_c}{q_a}\right)^c. \quad (\text{v})$$

In (v) above, we may write

$$\frac{p_b}{q_a} = p_{ba} \quad \text{and} \quad 1 - p_{ba} = q_{ba} = \frac{1 - p_a - p_b}{q_a} = \frac{p_c}{q_a}.$$

Since  $(r-a) = (b+c)$ , the general term of the binomial  $(q_{ba} + p_{ba})^{r-a}$  is

$$\frac{(r-a)!}{b! c!} p_{ba}^b \cdot q_{ba}^c = \frac{(r-a)!}{b! c!} \left(\frac{p_b}{q_a}\right)^b \left(\frac{p_c}{q_a}\right)^c.$$

Thus the second factor on the right of (v) is the  $(r-a)$ -fold sample weighted mean value of  $b^i$  when the unit sample expectation of extracting an item of class  $B$  is  $p_{ba}$ . It is therefore the  $i$ th zero moment of the distribution defined by successive terms of the expansion of  $(q_{ba} + p_{ba})^{r-a}$ .<sup>\*</sup> We may write it therefore as  $\mu_i(b_a)$ . In particular,

$$\mu_1(b_a) = (r-a)p_{ba};$$

$$\mu_2(b_a) = (r-a)p_{ba} q_{ba} + (r-a)^2 p_{ba}^2$$

$$= rp_{ba} q_{ba} - p_{ba} (q_{ba} + 2rp_{ba})a + p_{ba}^2 (a^2 + r^2).$$

<sup>\*</sup> In the symbolism of 11.01–11.04, used elsewhere in 12.03, the operation illustrated by (v) is equivalent to writing

$$\mu_{h \cdot i \cdot 0} = E(a^h b^i) = E_a[a^h \cdot E_{ba}(b^i)].$$

The operation  $E_{ba}(b^i)$  here signifies taking the mean of the  $i$ th power of  $b$  for a fixed value of  $a$ . Hence it is the  $i$ th moment of the  $(r-a)$  fold distribution for a residual universe in which the proportion of items of class  $B$  is  $p_b q_a^{-1}$ . It is evidently immaterial which way we write

$$E_a[a^h \cdot E_{ba}(b^i)] = \mu_{h \cdot i \cdot 0} = E_b[b^i \cdot E_{ab}(a^h)].$$



By substitution in (v) we thus obtain

$$\begin{aligned} E(a^2b^2) &= r^2p_{ba}^2 + rp_{ba}q_{ba} \sum_{r=0}^{r=a} a^2 \cdot r_{(a)}p_a^a q_a^{r-a} - p_{ba}(q_{ba} + 2rp_{ba}) \sum_{r=0}^{r=a} r_{(a)}p_a^a q_a^{r-a} \\ &\quad + p_{ba}^2 \sum_{r=0}^{r=a} a^4 r_{(a)} \cdot p_a^a q_a^{r-a} \\ &= r^2p_{ba}^2 + rp_{ba}q_{ba} \cdot \mu_2(a) - p_{ba}(q_{ba} + 2rp_{ba}) \cdot \mu_3(a) + p_{ba}^2 \cdot \mu_4(a). \end{aligned}$$

Alternatively we can express the general term of the binomial in the form

$$\frac{r!}{b!(r-b)!} p_b^b q_b^{r-b} \cdot \frac{(r-b)!}{a!c!} \left(\frac{p_a}{q_b}\right)^a \left(\frac{p_c}{q_b}\right)^c = r_{(b)}p_b^b q_b^{r-b} \cdot (r-b)_{(a)}p_a^a q_b^c.$$

Whence we may also write

$$E(a^2b^2) = r^2p_{ba}^2 + rp_{ab} \cdot q_{ab} \cdot \mu_2(b) - p_{ab}(q_{ab} + 2rp_{ab})\mu_3(b) + p_{ab}^2 \cdot \mu_4(b).$$

Hence we have

$$\begin{aligned} 2E(a^2b^2) &= (r^2p_{ba}^2 + rp_{ba}q_{ba})\mu_2(a) + (r^2p_{ab}^2 + rp_{ab}q_{ba})\mu_2(b) - p_{ba}(q_{ba} + 2rp_{ba})\mu_3(a) \\ &\quad - p_{ab}(q_{ab} + 2rp_{ab})\mu_3(b) + p_{ba}^2 \cdot \mu_4(a) + p_{ab}^2 \cdot \mu_4(b). \end{aligned}$$

In this expression

$$p_{ab} = \frac{p_a}{q_b}; \quad q_b = 1 - p_b; \quad q_{ab} = \frac{p_c}{q_b}.$$

Since  $M_a = rp_a$  and  $M_b = rp_b$ , we derive

$$\begin{aligned} \frac{2E(a^2b^2)}{M_a M_b} &= \frac{p_c + M_b}{M_a q_a^2} \mu_2(a) + \frac{p_c + M_a}{M_b q_b^2} \mu_2(b) - \frac{(p_c + 2M_b)}{r M_a q_a^2} \mu_3(a) \\ &\quad - \frac{(p_c + 2M_a)}{r M_b q_b^2} \mu_3(b) + \frac{M_b}{r^2 M_a q_a^2} \mu_4(a) + \frac{M_a}{r^2 M_b q_b^2} \mu_4(b). \end{aligned} \quad (vi)$$

The corresponding terms of the expressions involving  $E(a^2c^2)$  and  $E(b^2c^2)$  are definable by inspection, and (i) is now reducible to an expression involving the 2nd, 3rd and 4th zero moments of the  $A$ -score,  $B$ -score and  $C$ -score distributions. We may collect terms involving moments of the  $A$ -score distribution as follows:

$$\left[ \frac{p_c + M_b}{M_a q_a^2} + \frac{p_b + M_c}{M_a q_a^2} \right] \mu_2(a) - \left[ \frac{p_c + 2M_b}{r M_a q_a^2} + \frac{p_b + 2M_c}{r M_a q_a^2} \right] \mu_3(a) + \left[ \frac{M_b}{r^2 M_a q_a^2} + \frac{M_c}{r^2 M_a q_a^2} + \frac{1}{M_a^2} \right] \mu_4(a).$$

Since  $(p_b + p_c) = 1 - p_a = q_a$  and  $M_b + M_c = (r - M_a) = r q_a$  the above reduces to

$$\frac{r+1}{M_a q_a} \mu_2(a) - \frac{2r+1}{r M_a q_a} \mu_3(a) + \frac{1}{M_a^2 q_a} \mu_4(a).$$

We now recall the expressions for the 2nd, 3rd and 4th zero moments of the  $A$ -score distribution whose definitive binomial is  $(q_a + p_a)^r$ :

$$\begin{aligned} \mu_2(a) &= rp_a q_a + r^2 p_a^2 = M_a q_a + M_a^2; \\ \mu_3(a) &= rp_a q_a + 3r^2 p_a^2 q_a - 2r p_a^3 q_a + r^3 p_a^3 \\ &= M_a q_a + 3M_a^2 q_a - 2M_a p_a q_a + M_a^3; \\ \mu_4(a) &= rp_a q_a - 6r p_a^2 q_a^2 + 7r^2 p_a^2 q_a - 11r^2 p_a^3 q_a + 6r^3 p_a^3 q_a + r^4 p_a^4 \\ &= M_a q_a - 6M_a p_a q_a^2 + 7M_a^2 q_a - 11M_a^2 p_a q_a + 6M_a^3 q_a + M_a^4. \end{aligned}$$





The terms involving 4th moments reduce to

$$\begin{aligned} \frac{1}{M_a} + \frac{1}{M_b} + \frac{1}{M_c} - \frac{12}{r} + 21 - 11(p_a + p_b + p_c) + 6(M_a + M_b + M_c) + \frac{M_a^2}{q_a} + \frac{M_b^2}{q_b} + \frac{M_c^2}{q_c} \\ = 10 + 6r - \frac{12}{r} - 4r^2 + r^2 \left( \frac{1}{q_a} + \frac{1}{q_b} + \frac{1}{q_c} \right) + \left( \frac{1}{M_a} + \frac{1}{M_b} + \frac{1}{M_c} \right) \quad (\text{xiv}) \end{aligned}$$

By substitution of (xi)–(xiv) in (x) we have

$$\mu_2(C_3) = 8 - \frac{13}{r} + \left( \frac{1}{M_a} + \frac{1}{M_b} + \frac{1}{M_c} \right) \quad . \quad . \quad . \quad . \quad (\text{xv})$$

Evidently, the expression on the right approaches the limit 8 in agreement with (ii) if  $r$  is very large. In the same way, we may show that

$$\mu_3(C_3) \simeq 48 = f(f+2)(f+4) \quad \text{and} \quad \mu_4(C_3) \simeq 384 = f(f+2)(f+4)(f+6)$$

on the same assumption. However, it is equally evident that the moments of the statistic under discussion do *not* closely agree with those of the exact Chi-Square distribution unless the size of the sample is in fact very large. The closeness of the approximation and the sign of error depend not only on  $r$  but on the expected class proportions.

The exact definition of moments of higher order than  $\mu_2$  for the 3-class case introduces no new matter of principle. It will therefore suffice to cite the results, *viz.* :

$$\mu_3(C_3) = 48 - \frac{4(79r - 69)}{r^2} + \frac{(28r - 31)}{r} \sum_{i=1}^3 \frac{1}{M_i} + \sum_{i=1}^3 \frac{1}{M_i^2} \quad . \quad . \quad . \quad . \quad (\text{xvi})$$

$$\begin{aligned} \mu_4(C_3) = 384 - \frac{(6368r^2 - 18123r + 12150)}{r^3} + \frac{2(340r^2 - 1109r + 862)}{r^2} \sum_{i=1}^3 \frac{1}{M_i} \\ + \frac{3(r - 1)}{r} \left( \sum_{i=1}^3 \frac{1}{M_i} \right)^2 + \frac{4(30r - 31)}{r} \sum_{i=1}^3 \frac{1}{M_i^2} + \sum_{i=1}^3 \frac{1}{M_i^3} \quad . \quad . \quad . \quad . \quad (\text{xvii}) \end{aligned}$$

It is easy, if also laborious, to recognise the common pattern for  $C_2$ ,  $C_3$ , etc. Thus more generally for a system of  $n = (f + 1)$  classes the moments approach those of the Chi-Square variate for  $f$  degrees of freedom, e.g.

$$\mu_2(C_n) = (n - 1)(n + 1) - \frac{n^2 + 2n - 2}{r} + \sum_{i=1}^n \frac{1}{M_i} \quad . \quad . \quad . \quad . \quad . \quad . \quad (\text{xviii})$$

$$\begin{aligned} \mu_3(C_n) = (n - 1)(n + 1)(n + 3) - \frac{r(3n^3 + 21n^2 + 24n - 26) - 2(n + 3)(n^2 + 6n - 4)}{r^2} \\ + \frac{r(3n + 19) - (3n + 22)}{r} \sum_{i=1}^n \frac{1}{M_i} + \sum_{i=1}^n \frac{1}{M_i^2} \quad . \quad . \quad . \quad (\text{xix}) \end{aligned}$$

The accompanying tables (1-3) illustrate the exactitude of (xv)–(xvii) for the 3-class case,  $Q$  being there the sample value of the Chi-Square statistic and  $f$  the frequency of a sample of specified structure. Thus the column totals for  $fQ$ ,  $fQ^2$ ,  $fQ^3$ ,  $fQ^4$  are the numerical values of  $\mu_1$ ,  $\mu_2$ ,  $\mu_3$ ,  $\mu_4$ . The reader may still ask : how close must be the correspondence between the moments of the statistic under consideration and those of the Chi-Square distribution of

15.02, if we are to use the latter legitimately as a fitting curve? The question invites laborious and formidable computations. It admits of no complete answer to date, and is worthy of exploration with the aid of the newest electronic machines.

TABLE 1

Three-Class Universe ( $p_a = \frac{1}{4}$ ,  $p_b = \frac{1}{3}$ ,  $p_c = \frac{5}{12}$ )

$$Q = \frac{(A - M_a)^2}{M_a} + \frac{(B - M_b)^2}{M_b} + \frac{(C - M_c)^2}{M_c}$$

(a) Unit Sample.

Sample Structure			Frequency					
A	B	C	(f)	Q	fQ	fQ <sup>2</sup>	fQ <sup>3</sup>	fQ <sup>4</sup>
1	0	0	$p_a = \frac{1}{4}$	$\frac{(1 - \frac{1}{4})^2}{\frac{1}{4}} + \frac{(-\frac{1}{3})^2}{\frac{1}{3}} + \frac{(-\frac{5}{12})^2}{\frac{5}{12}} = 3$	$\frac{3}{4}$	$\frac{9}{4}$	$\frac{27}{4}$	$\frac{81}{4}$
0	1	0	$p_b = \frac{1}{3}$	$\frac{(-\frac{1}{4})^2}{\frac{1}{4}} + \frac{(1 - \frac{1}{3})^2}{\frac{1}{3}} + \frac{(-\frac{5}{12})^2}{\frac{5}{12}} = 2$	$\frac{2}{3}$	$\frac{4}{3}$	$\frac{8}{3}$	$\frac{16}{3}$
0	0	1	$p_c = \frac{5}{12}$	$\frac{(-\frac{1}{4})^2}{\frac{1}{4}} + \frac{(-\frac{1}{3})^2}{\frac{1}{3}} + \frac{(1 - \frac{5}{12})^2}{\frac{5}{12}} = \frac{7}{5}$	$\frac{7}{12}$	$\frac{49}{60}$	$\frac{343}{300}$	$\frac{2401}{1560}$
Totals					2	$\frac{22}{5}$	$\frac{264}{25}$	$\frac{3398}{125}$

TABLE 2

Three-Class Universe ( $p_a = \frac{1}{4}$ ,  $p_b = \frac{1}{3}$ ,  $p_c = \frac{5}{12}$ ).

$$Q = \frac{(A - M_a)^2}{M_a} + \frac{(B - M_b)^2}{M_b} + \frac{(C - M_c)^2}{M_c}$$

(b) 2-fold Sample.

Sample Structure			f	Q	fQ	fQ <sup>2</sup>	fQ <sup>3</sup>	fQ <sup>4</sup>
A	B	C						
2	0	0	$p_a = \frac{1}{16}$	$\frac{(2 - \frac{1}{2})^2}{\frac{1}{2}} + \frac{(-\frac{2}{3})^2}{\frac{2}{3}} + \frac{(-\frac{5}{6})^2}{\frac{5}{6}} = 6$	$\frac{3}{8}$	$\frac{9}{4}$	$\frac{27}{2}$	81
1	1	0	$2p_ap_b = \frac{1}{6}$	$\frac{(1 - \frac{1}{2})^2}{\frac{1}{2}} + \frac{(1 - \frac{2}{3})^2}{\frac{2}{3}} + \frac{(-\frac{5}{6})^2}{\frac{5}{6}} = \frac{3}{2}$	$\frac{1}{4}$	$\frac{3}{8}$	$\frac{9}{16}$	$\frac{27}{32}$
1	0	1	$2p_ap_c = \frac{5}{24}$	$\frac{(1 - \frac{1}{2})^2}{\frac{1}{2}} + \frac{(-\frac{2}{3})^2}{\frac{2}{3}} + \frac{(1 - \frac{5}{6})^2}{\frac{5}{6}} = \frac{6}{5}$	$\frac{1}{4}$	$\frac{3}{10}$	$\frac{9}{25}$	$\frac{54}{125}$
0	2	0	$p_b^2 = \frac{1}{9}$	$\frac{(-\frac{1}{2})^2}{\frac{1}{2}} + \frac{(2 - \frac{2}{3})^2}{\frac{2}{3}} + \frac{(-\frac{5}{6})^2}{\frac{5}{6}} = 4$	$\frac{4}{9}$	$\frac{16}{9}$	$\frac{64}{9}$	$\frac{256}{9}$
0	1	1	$2p_bp_c = \frac{5}{18}$	$\frac{(-\frac{1}{2})^2}{\frac{1}{2}} + \frac{(1 - \frac{2}{3})^2}{\frac{2}{3}} + \frac{(1 - \frac{5}{6})^2}{\frac{5}{6}} = \frac{7}{10}$	$\frac{7}{36}$	$\frac{49}{360}$	$\frac{343}{3600}$	$\frac{2401}{36000}$
0	0	2	$p_c^2 = \frac{25}{144}$	$\frac{(-\frac{1}{2})^2}{\frac{1}{2}} + \frac{(-\frac{2}{3})^2}{\frac{2}{3}} + \frac{(2 - \frac{5}{6})^2}{\frac{5}{6}} = \frac{14}{5}$	$\frac{35}{72}$	$\frac{49}{36}$	$\frac{343}{90}$	$\frac{2401}{225}$
Totals					2	$\frac{31}{5}$	$\frac{638}{25}$	$\frac{60729}{500}$



TABLE 3

*Three-Class Universe* ( $p_a = \frac{1}{4}$ ,  $p_b = \frac{1}{3}$ ,  $p_c = \frac{5}{12}$ )

$$Q = \frac{(A - M_a)^2}{M_a} + \frac{(B - M_b)^2}{M_b} + \frac{(C - M_c)^2}{M_c}$$

*(c) 3-fold Sample.*

Sample Structure								
<i>A</i>	<i>B</i>	<i>C</i>	<i>f</i>	<i>Q</i>	<i>fQ</i>	<i>fQ</i> <sup>2</sup>	<i>fQ</i> <sup>3</sup>	<i>fQ</i> <sup>4</sup>
3	0	0	$p_a^3 = \frac{1}{64}$	9	$\frac{9}{64}$	$\frac{81}{64}$	$\frac{729}{64}$	$\frac{6561}{64}$
2	1	0	$3p_a^2p_b = \frac{1}{16}$	$\frac{10}{3}$	$\frac{5}{24}$	$\frac{25}{36}$	$\frac{125}{54}$	$\frac{625}{81}$
2	0	1	$3p_a^2p_c = \frac{5}{64}$	$\frac{47}{15}$	$\frac{47}{192}$	$\frac{2209}{2880}$	$\frac{103,823}{43,000}$	$\frac{4,879,681}{648,000}$
1	1	1	$6p_ap_bp_c = \frac{5}{24}$	$\frac{2}{15}$	$\frac{1}{36}$	$\frac{1}{270}$	$\frac{1}{2025}$	$\frac{2}{30,375}$
1	2	0	$3p_ap_b^2 = \frac{1}{12}$	$\frac{7}{3}$	$\frac{7}{36}$	$\frac{49}{108}$	$\frac{343}{324}$	$\frac{2401}{972}$
1	0	2	$3p_ap_c^2 = \frac{25}{192}$	$\frac{23}{15}$	$\frac{115}{576}$	$\frac{529}{1728}$	$\frac{12,167}{25,920}$	$\frac{279,841}{388,800}$
0	3	0	$p_b^3 = \frac{1}{27}$	6	$\frac{2}{9}$	$\frac{4}{3}$	8	48
0	2	1	$3p_b^2p_c = \frac{5}{36}$	$\frac{9}{5}$	$\frac{1}{4}$	$\frac{9}{20}$	$\frac{81}{100}$	$\frac{729}{500}$
0	1	2	$3p_bp_c^2 = \frac{25}{144}$	$\frac{6}{5}$	$\frac{5}{24}$	$\frac{1}{4}$	$\frac{3}{10}$	$\frac{18}{50}$
0	0	3	$p_c^3 = \frac{125}{1728}$	$\frac{21}{5}$	$\frac{175}{576}$	$\frac{245}{192}$	$\frac{343}{64}$	$\frac{7203}{320}$
Totals					2	$\frac{34}{5}$	$\frac{7227}{225}$	$\frac{652,318}{3375}$

## CHAPTER 20

# SECOND THOUGHTS ON SIGNIFICANCE\*

### 20.00 STATISTICAL INFERENCE

MUCH of the content of this volume deals with test procedure designed to assess the credentials of a unique null hypothesis. It would be unfitting to conclude it without reference to growing uneasiness with respect to the role of the unique null hypothesis in statistical reasoning. Indeed, it is a remarkable circumstance that our own generation has simultaneously experienced unprecedented eagerness to exploit new tests of significance and vigorous controversy on a wide front with respect to the credentials of statistical inference. On all sides we hear that statistical theory is the logic of the sciences ; but at least three divergent views about its rationale are current and have advocates of no mean intellectual stature. Meanwhile, it is not feasible to offer a definition of statistical inference acceptable to all mathematicians who concern themselves with the theory of probability or to all practical statisticians. Any such definition presupposes an answer to the age-old question : what is truth ? Any answer to the latter presupposes a personal credo embodying the relation of human knowledge to the external world. That of the writer is in the broadest sense of the term behaviourist. Accordingly, we shall here assume that (a) any recipes for arriving at truth (*rules of inference*) on the basis of inescapably imperfect acquaintance with the real world have as the end in view an unequivocal assertion coupled with an admission of liability to error ; (b) what distinguishes the recipes we call statistical inference is that this admission—the *uncertainty safeguard* of the assertion—is *numerically specifiable within an assumed framework of indefinitely protracted repetition*. Thus the uncertainty safeguard is the probability of false statement.

From a practical viewpoint, it is useful to distinguish sharply between two techniques of statistical inference, though we shall later seek for a formula embracing both :

- (a) *test procedures*, including the traditional null hypothesis significance tests, ostensibly devised to adjudicate on the merits of particular hypotheses ;
- (b) *methods of estimation*, the aim of which is to make legitimate statements about numerical characteristics of a universe or subuniverse on the basis of information supplied by a sample.

Within the domain of test decisions,† it is essential to distinguish between different targets of statistical inference :

- (i) to decide whether to regard a particular hypothesis as true or false ;
- (ii) to limit the risk of rejecting it if it is indeed correct.

\* I am greatly indebted to Raymond Wrighton for many (to me) profitable discussions of issues raised in this chapter which incorporates the substance of joint papers (Hogben and Wrighton) on The Statistical Theory of Therapeutic and Prophylactic Trials in the *British Journal of Social Medicine* (1952).

† F. J. Anscombe (1951), *Mind*, Vol. 60, makes a three-fold distinction :

“ It is worthwhile to distinguish different purposes one may have in accepting a hypothesis : (i) to base an administrative decision on, (ii) for further testing and confirmation, (iii) for acceptance into the corpus of scientific knowledge, to be relied on in future work. There are risks, variously assessable, in coming to decisions in all three cases. For example, in case (iii), if the hypothesis is later found to be seriously false a lot of effort in investigating other points may have been wasted. Just as with prior confidences, risks are rather vague in magnitude, but in a formal theory it would be tempting to postulate a complete numerical risk-function.”



Broadly speaking, this dichotomy tallies with a useful distinction between two types of statistical inference definable as follows:

- (i) *unconditional*, if the uncertainty safeguard specifies the unconditional probability of the falsity of the assertion itself;
- (ii) *conditional*, if the uncertainty safeguard merely specifies the probability of rejecting the relevant hypothesis when it is true.

In symbolic form we may express the unconditional probability of false assertion as  $P_f = (1 - P_t)$ , and the conditional probability of false assertion within the framework of a particular hypothesis  $A$  as  $P_{f.a} = (1 - P_{t.a})$ . Thus an assertion of the form  $P_t = x$  is an example of unconditional statistical inference. Besides these two forms of statement we may make one of the form  $P_t \geq x$ . Evidently the more exact assertion  $P_t = 0.95$  has no pragmatic priority over the less definite assertion  $P_t \geq 0.95$ ; and we may prefer to regard a statement expressed in the form  $P_t \geq x$  as an example of (i) if we deem  $(1 - x)$  to be an *acceptable* level of uncertainty. On the other hand, it serves no useful purpose to make an assertion of the form  $P_t \geq 0.30$  if we regard any figure above 5 per cent. as an unacceptable level of uncertainty. Hence we shall have no practical interest in stating an inference of the form  $P_t \geq x$  unless we should be content with the assertion  $P_t = x$ . Otherwise any useful statement of statistical inference we may undertake conforms to (ii).

*First Aid for Inequalities.* In what follows we shall make more use than heretofore of inequalities referred to briefly on p. 10 of Vol. I. In higher school and elementary college courses one deals mainly with equations. One has therefore little experience of the importance of, or opportunity to get familiar with, statements involving the ideograms  $<$  or  $>$  and  $\leq$  (*not greater than*) or  $\geq$  (*not less than*). The student of statistical theory should be thoroughly familiar with their use. Here follows a short dictionary of meanings which the reader may interpret or test by substituting whole numbers for literal symbols.

- (i)  $m \geq x \geq k$  or  $k \leq x \leq m$  means:  $x$  lies in the range  $k$  to  $m$  inclusive.
- (ii)  $m \geq x > k$  or  $k < x \leq m$  means:  $x$  is greater than  $k$  and no greater than  $m$ .
- (iii)  $m > x \geq k$  or  $k \leq x < m$  means:  $x$  is less than  $m$  and not less than  $k$ .
- (iv) the two statements  $x > m$  and  $x \leq m$  constitute an exclusive binary classification of the range of values  $x$  may assume, as do also the two statements  $x \geq m$  and  $x < m$ .
- (v) any of the following statements constitute an exclusive three-fold split of the range in which  $x$  may lie:

$$(a) \ x > k; \ k \geq x \geq m; \ x < m$$

$$(b) \ x \geq k; \ k > x \geq m; \ x < m$$

$$(c) \ x > k; \ k \geq x > m; \ x \leq m$$

$$(d) \ x \geq k; \ k > x > m; \ x \leq m$$

- (vi) The following rules of sign reversal are important:

$$k - b < k - a \text{ when } b > a$$

$$k - b \leq k - a \text{ when } b \geq a$$

(vii) If we denote the probability that the score  $x$  is no greater than  $m$  by  $P(x \leq m)$  and probabilities referable to other statements about the interval in which  $x$  lies in accordance with the same pattern, certain important identities derive from the addition theorem in virtue of the above, e.g.

$$P(x \geq m) + P(x < m) = 1$$

$$P(x \geq m) > 1 - \alpha \text{ if } P(x < m) < \alpha$$

## 20.01 STATISTICAL INSPECTION

In the foregoing discussion we have made a provisionally clear-cut distinction between estimation and test decision ; but we shall later see that the views of a growing and influential school of theoretical statistics make it possible to formulate a definition of statistical inference which exhibits the procedure of test decisions as a limiting case of the procedure of estimation. A new orientation, of which the theory is due to J. Neyman, E. S. Pearson and A. Wald, registers the impact on statistical theory of practices and metaphors which have developed over a long period under a shroud of trade secrecy within the research laboratories of large corporations, more especially the Bell Telephone Company. Much that is otherwise mysterious and highly abstract comes to life against the background of industrial practice. We shall therefore be better able to appreciate a fresh approach to the problem of significance, if we acquaint ourselves with some elements of the technique of statistical inspection (*quality control*) in commerce or industry. To do so some new terms will be necessary. On that account we must now digress.

At the most elementary (*p*-chart) level the aim of statistical inspection is to ensure that the production process is working to schedule. The assumption is that no machine is perfect in the sense that the output is of uniform excellence. All we can hope for is that it continues to deliver samples of products with a fixed and satisfactory mean value in accordance with a known law of error, e.g. normal or binomial. The scoring system may be representative (e.g. duration of life of an electric light bulb) or taxonomic (e.g. percentage of inactive ampoules). If the sample score lies outside a range deemed satisfactory (e.g.  $3\sigma$  level), the inspection system recommends to the management overhaul of the machinery to ascertain whether the result is a fluke. Otherwise we may speak of the process as being in *statistical equilibrium* (Fig. 127).

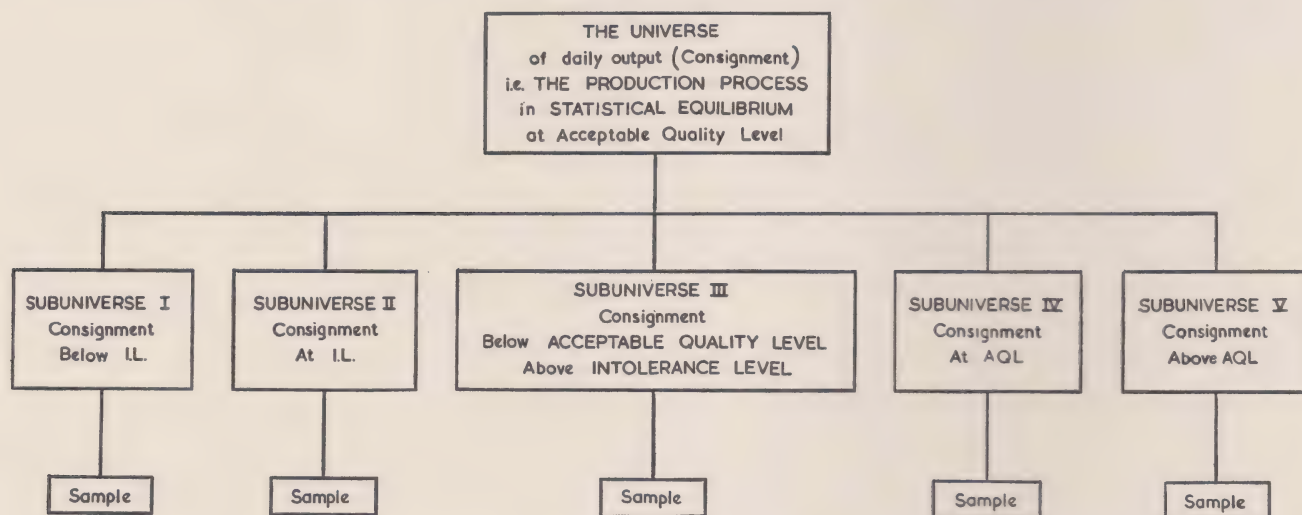


FIG. 127. Sampling in the Stratified Universe of a Production Process in statistical equilibrium.

In what follows we assume that this is so. The mean daily output will then be up to standard. That some consignments will be below it (Fig. 128) is then fully consistent with the possibility that the process is working as well as may be.

Thus the manufacturer or seller can at most guarantee that the product will very rarely fall short of a standard of precision called the *acceptable quality level* (a.q.l.), e.g. that the sectional area of two by two inch wooden battens will not be less than 3.9 sq. in. or that the proportion of inert ampoules of post-pituitary extract in a consignment will not exceed one per cent. Complete



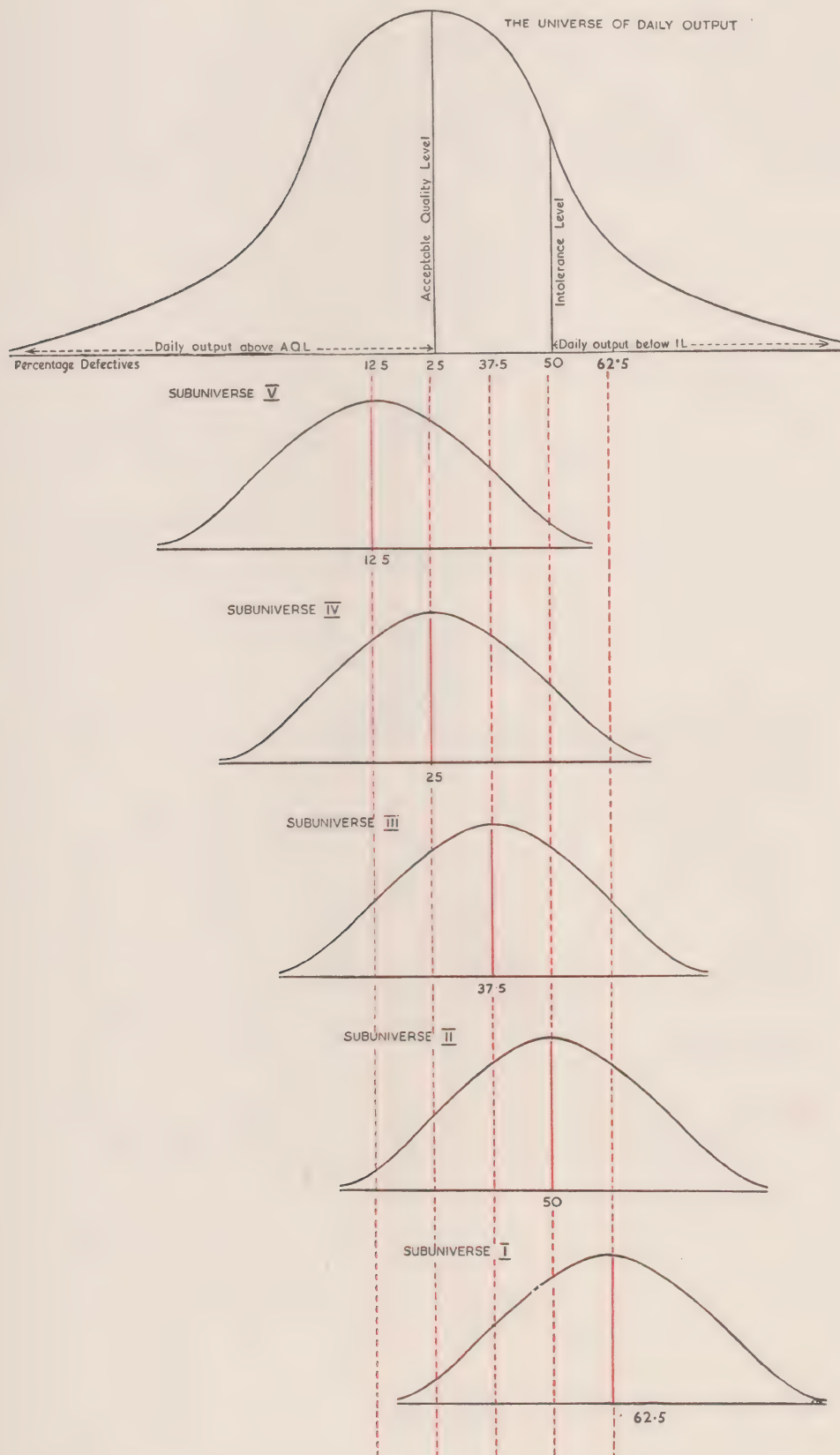


FIG. 128. Sample Distributions for subuniverses of Fig. 127.

inspection to maintain this standard would be costly if practicable. Often it is impracticable, because testing the product (e.g. mean duration of life of an electric light bulb or activity of a glandular extract) involves destroying it. What we call quality control is therefore a sampling procedure. Evidently, a sample from a consignment above a.q.l. might be below it. Thus a 500-fold sample at random from a consignment of ampoules of which only 0.9 per cent. are defective might well contain 6 defectives (1.2 per cent.). To reject every consignment as below the 1 per cent. guaranteed a.q.l. if the sample contained more than 1 per cent. defectives

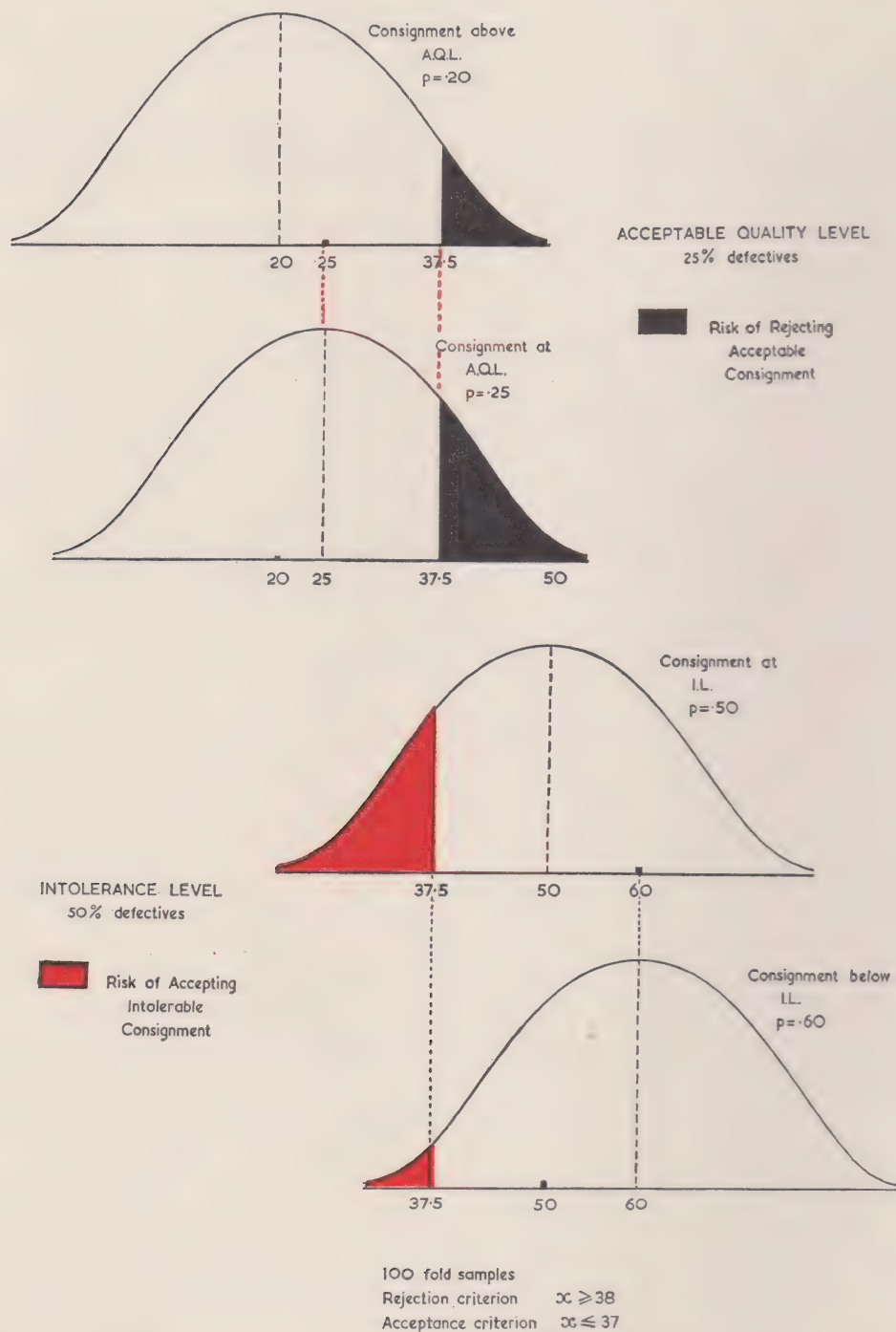


FIG. 129. The risk of rejecting a consignment above Acceptable Quality Level is less than the risk of rejecting a consignment at a.q.l. and the risk of accepting a consignment below Intolerance Level is less than the risk of accepting it at i.l.



would therefore be wasteful. So one aim of quality control is to ensure that the producer will rarely discard a consignment, if it is *at or above* a.q.l.

Given the law of distribution, e.g. a normally distributed sample mean or a binomial distribution of proportionate defectives, the producer or seller may decide to *reject* consignments on the basis of sample structure in such a way as to exclude in the long run only  $\alpha$  per cent. (e.g. 5 per cent.) of those at exactly a.q.l. We call this the *producer's risk*. This procedure, which ensures a risk less than  $\alpha$  per cent. of rejecting consignments above a.q.l., is consistent with endorsing some consignments which do *not* in fact satisfy the a.q.l. guaranteed to the consumer. To safeguard the confidence of the latter, the seller or producer must ensure a low risk of releasing consignments too defective to be tolerable. This presupposes some standard we shall here call the *level of intolerance* (i.l.) and a criterion of acceptance comparable with the criterion of rejection, i.e. acceptance on condition that the risk of release at i.l. is only  $\beta$  per cent. (e.g. 5 per cent.). We call this *consumer's risk*.\* The risk of rejecting a consignment *above* a.q.l. is less than producer's risk as already defined, i.e. risk of rejection *at* a.q.l. The risk of accepting a consignment *below* i.l. is likewise less than the so-called consumer's risk, i.e. risk of rejecting *at* i.l. (Fig. 129).

To give a more precise meaning to these terms it is essential to be clear that they refer in this context to *opposite* tails of a sample distribution. Accordingly, we must first recall (Vol. I, Chapter 5) the distinction between *vector* and *modular* assessment of risk. Thus the modular risk that a sample value of a normally distributed variate will deviate from the true mean by more than  $\pm 1.96\sigma$  is 5 per cent.; but the 5 per cent. level is at  $-1.64\sigma$  for the risk that a score will fall short of the mean and at  $+1.64\sigma$  for the risk that it will exceed the mean by so much. A fictitious example will clarify the issue. We shall suppose that :

- (a) the producer sets his level of acceptability for a consignment of battens at a mean figure  $M_a = 20$  mm. in thickness and the level of intolerance at a mean figure  $M_b = 18$  mm. in thickness ;
- (b) under ascertained working conditions of the sawmill the standard error of a 100-fold sample mean is 0.75 mm. with an approximately normal distribution.

If inspection shows that the sample mean of a 100-fold sample of a particular load is  $\bar{x} = 18.75$ , the equivalent standard scores will be :

- (a) if from a consignment at a.q.l.

$$\frac{18.75 - 20}{0.75} = -1.66;$$

- (b) if from a consignment at i.l.

$$\frac{18.75 - 18}{0.75} = 1.$$

The decision to accept only samples above 18.75 would thus involve a producer's risk at the  $1.66\sigma$  level and a consumer's risk at the  $1\sigma$  level. From the table of the normal we find that the area up to  $-1.66\sigma$  is 0.049 (nearly 5 per cent.) and the area beyond  $+\sigma$  is 0.159 (nearly 16 per cent.). If the true mean were above the a.q.l. the corresponding standard score would be numerically greater than  $-1.66$  and of the same sign. If below i.l. it would be numerically greater than  $+1$  and of the same sign. If our criterion of acceptability is 18.75 for the 100-fold

\* The expression *consumer's risk* has a taint of uplift and is somewhat misleading on that account. It suggests that the primary end in view is to look after the interests of the consumer. It is more precise to regard the end in view as that of limiting the producer's risk of losing the consumer's goodwill.

sample mean, the risk of rejecting a consignment at or above a.q.l. is thus less than 5 per cent. and the risk of accepting a consignment at or below i.l. is 16 per cent. or less.

Suppose now that we equalise the two risks, i.e. choose a score criterion ( $x$ ) which makes the two standard scores numerically equal with opposite signs, so that

$$\frac{x - 18}{0.75} = \frac{20 - x}{0.75}.$$

Thus  $x = 19$ . In this event, the numerical value of the standard score is 1.3 and the table of the normal integral gives the area of the excluded tails as 0.092 or about 9 per cent. risk for both consumer and producer. By making the size ( $r$ ) of sample larger we can, of course, make the variance of the sample mean smaller and the standard score itself larger. The s.d. of the two distributions will be in the ratio  $r^{-\frac{1}{2}} : 100^{-\frac{1}{2}}$ . In any case, the two risks will be equal if  $x = 19$ . If we want to keep the two risks at 5 per cent. or  $1.64\sigma$  level we therefore have :

$$\frac{\sqrt{r}}{10} \cdot \frac{20-19}{0.75} = 1.64 = \frac{-(18-19)}{0.75} \cdot \frac{\sqrt{r}}{10},$$

$$\therefore r \simeq 151.$$

As an alternative illustration we may (again fictitiously) suppose that the pharmacist sets an a.q.l. for the proportion of below-standard ampoules of a preparation at 25 per cent. and an intolerance level of 50 per cent. For small samples (under 50) the normal approximation will be poor and for even larger samples the half interval correction (p. 116) will make a big difference to our assessment of per cent. risk. For illustrative purposes, we may therefore content ourselves with defining the risk in terms of the critical ratio ( $h$ ). First suppose that  $r = 27$  (Fig. 130). We shall denote by  $x$  the number of defectives in the 27-fold sample and define  $x_c$  so that we : (a) reject a consignment if  $x > x_c$ ; (b) accept a consignment if  $x < x_c$ . We now express  $x_c$  in terms of the true mean and the sample s.d. If  $p_1 = \frac{1}{4}$  (a.q.l. 25 per cent.) :

$$x_c = rp_1 + h\sqrt{rp_1q_1} = \frac{27}{4} + \frac{h\sqrt{(81)}}{4} = \frac{27 + 9h}{4}.$$

If  $p_2 = \frac{1}{2}$  (i.l. 50 per cent.), we have likewise :

$$x_c = \frac{27}{2} - \frac{h\sqrt{27}}{2} = \frac{27 - 3h\sqrt{3}}{2}.$$

Whence we have

$$\frac{27 + 9h}{4} = \frac{54 - 6h\sqrt{3}}{4}.$$

In this case  $h$  is numerically a little less than 1.4 and  $x_c \simeq 9.8$ , i.e. we should reject consignments if the 27-fold sample contained 10 or more defectives and accept them if they contained 9 or less.

Now we may wish to make our risk smaller, let us say at  $2\sigma$  level. If so, we reject when

$$x > \frac{27}{4} + \frac{2\sqrt{81}}{4} \quad \text{i.e. } x = 12 \text{ or over.}$$

We should accept if

$$x < \frac{27}{2} - 3\sqrt{3} \quad \text{i.e. } x = 8 \text{ or less.}$$

This would leave us with consignments about which we could make no decision to cover both risks on equal terms (at the  $2\sigma$  level), i.e. if the 27-fold sample yields 9, 10 or 11 defectives. We



## INCOMPLETE INSPECTION PLAN TO SAFEGUARD PRODUCERS' RISK AND CONSUMERS' RISK

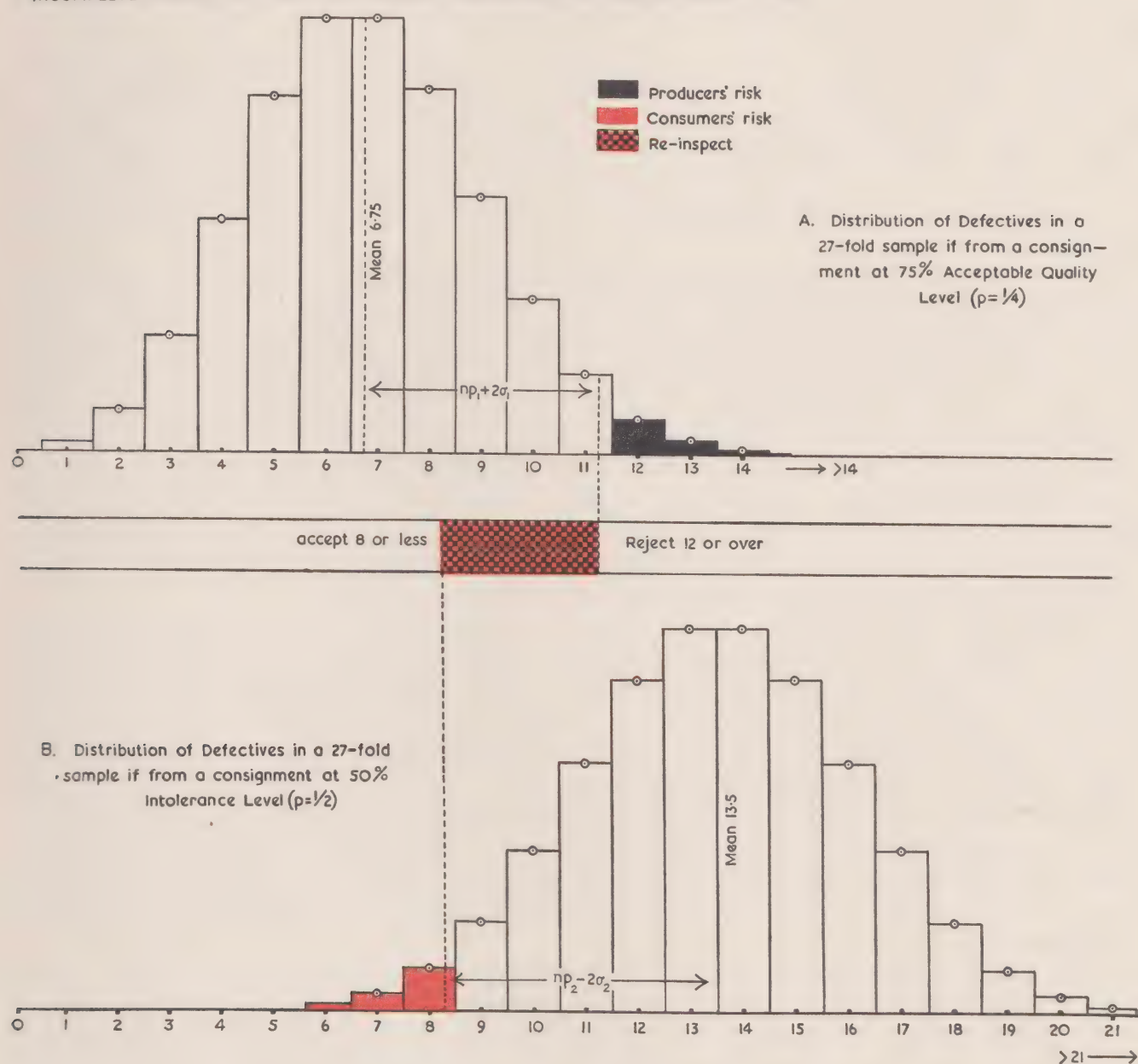


FIG. 130. Sampling on the basis of an incompletely decisive rejection-acceptance criterion. Here  $p$  is the proportion of defectives. The values  $p = 0.25$  (75 per cent. up to standard) and  $p = 0.50$  (only 50 per cent. up to standard) respectively define a.q.l. and i.l.

could, of course, retest by taking another 27-fold sample ; but the outcome might also be inconclusive. So it is more economical to cut our cloth to the standard set by deciding in advance what size ( $r$ ) of sample will make  $h = 2$  (or other prescribed risk criterion) when

$$rp_1 + h\sigma_1 = x_c = rp_2 - h\sigma_2.$$

When  $h = 2$  we find that  $r \simeq 56$ , in which case we should reject consignments if the 56-fold sample yielded 21 or more defectives, and pass them if it yielded 20 or less.

At this point, it is important to realise that we cannot have the best of both worlds by adding the result obtained from a first and inconclusive trial based on a 27-fold sample to that of a second sample of 29 in order to make up a 56-fold trial which must give a conclusive result.

This procedure signifies that we do not give a subsequent 29-fold sample the opportunity to pair off with a 27-fold sample unless the latter contains 9-11 defectives. Thus only 3 out of 28 classes of 27-fold sample would go to the making of the sample of 56; and we can make up a sample of 56 in accordance with the assumption of random choice by adding samples of 27 and 29 only if every possible sample of 27 has an equal chance of association with every possible sample of 29. This raises the question: is it necessary to prescribe in advance a sample large enough to ensure a conclusive result? In other words, can we devise an admissible inspection scheme which we can terminate so soon as the result is decisive? We shall return to this issue at a later stage.

Here it will be well to emphasise that the foregoing outline of the inspection problem is highly simplified for expository use in connexion with the main theme of this chapter. We have used highly fictitious examples for arithmetical simplicity. We have assumed: (a) as will usually be true, the critical score  $x_c$  is not an integer, so that all scores obtainable lie on one side of it or the other; (b) since the sample is a small fraction of a large consignment, we can dispense with the replacement condition; (c) we know the true variance of the sample distribution when we score by the representative method and have good enough reason for regarding the normal approximation as adequate. Actually, a normal approximation will not be a good one when the method of scoring is taxonomic, as in the last example. Commonly, the admissible proportion of defectives will be very much smaller than 0.25 or 0.5 and a Poisson distribution might then give us a better picture of what we are doing.

Within the same framework of limitations let us now look at our last illustration from a different point of view. We have chosen to adopt as our criterion of rejection for the 56-fold sample  $x \geq 22$  and as our criterion of acceptance  $x \leq 21$ . This ensures a risk of approximately 5 per cent. that we shall reject consignments at a.q.l. ( $p = 0.25$ ) and an equivalent risk that we shall accept consignments at i.l. ( $p = 0.5$ ), i.e. we shall accept 95 per cent. or more consignments if at or above a.q.l. and reject 95 per cent. or more if at or below i.l. It goes without saying that the quality of some consignments will neither be up to a.q.l. or as low as i.l. For instance, the true proportion of defectives in the consignment might be 0.38, in which event the 56-fold sample mean would lie between 21 and 22. Our rejection-acceptance criterion would then ensure about 50 per cent. risk of rejecting and the same risk of accepting such a consignment.

We speak of the inspection plan as complete if it always leads to a decision at both a prescribed producer's risk ( $\alpha$ ) and a prescribed consumer's risk ( $\beta$ ). We then have a model of what Wald calls a *decision*, in contradistinction to a *significance*, test. Formally we may describe the plan in terms of a rule to reject one or other alternative hypothesis: hypothesis  $A$  that the consignment is at or above acceptable quality level, i.e.  $M \geq M_a$ ; hypothesis  $B$  that the consignment is at or below intolerance level, i.e.  $M \leq M_b$ . The test, i.e. the inspection plan itself, is the rule: reject the consignment only if the sample score  $x < x_c$ . In effect, we therefore say: reject hypothesis  $A$  if  $x < x_c$ , and reject hypothesis  $B$  if  $x \geq x_c$ . Our decision is verbally equivalent *either* to denying that the consignment is up to a.q.l. *or* to denying that the consignment is at or below i.l. Neither decision implies the denial of the possibility that the consignment mean ( $M$ ) lies between the two levels ( $M_a < M < M_b$ ).

We chose  $x_c$  so that  $\alpha$  is the probability of rejecting the consignment at a.q.l. ( $M = M_a$ ) and  $\beta$  is that of accepting one at i.l. ( $M = M_b$ ). Since the probability of rejecting a consignment above a.q.l. will be less than  $\alpha$ , and that of accepting one below i.l. will be less than  $\beta$ , we may say that the risk of rejecting hypothesis  $A$  ( $M < M_a$ ) when it is true is  $P_{f.a} \leq \alpha$  and the risk of rejecting hypothesis  $B$  ( $M \leq M_b$ ) when it is true is  $P_{f.b} \leq \beta$ . To choose  $x_c$ , our rejection score criterion, so that the risks  $\alpha$  and  $\beta$  are themselves acceptable (e.g.  $\alpha = 0.05 = \beta$ ) we must



prescribe the size of the sample in advance. Our inspection plan which guarantees a decision to act in one way or the other is thus a test which guarantees minimum risk of wrongly rejecting one or other prescribed hypothesis, *if we consistently follow the same rejection rule for samples of the same size*. How such a prescription of test procedure differs from that of Yule and of Fisher will be the theme of comment at a later stage.

## 20.02 BAYES' THEOREM AND THE SEQUENTIAL RATIO

In Chapter 5 of Vol. I we have exhibited a much-discussed theorem of *unconditional* statistical inference, that of Thomas Bayes (1763), as a balance sheet which sets out what information we require in order to specify the long-run frequencies with which we shall arbitrate correctly on the assumption that one or other of an exhaustive set of hypotheses is correct. We conceive each hypothesis as the assertion that a sample of specified composition comes from a sub-universe also of specified composition, leaving open the possibility that the composition of different sub-universes may be identical. Thus the sub-universes may be urns containing coloured balls. We shall then speak of them as urns of the same type if they contain balls of the same colour in the same proportions. This definition suffices if we sample with replacement. If we sample *without* replacement, we must make the additional assumption that urns of the same type contain the same total number of balls. The most general specification of the sample taken from one or other urn will, of course, be the term of a multinomial expansion; but it will here suffice to specify the sample on the assumption that each sub-universe is of the 2-class kind, as when we distinguish balls by colour as *red* and *other*. For simplicity, we shall also assume (unless otherwise stated) that sampling is *with* replacement.

As an example of the stratified universe of Bayes we may envisage a set of 10 urns constituted as follows:

					No. of Urns	Proportion of Red balls
Type I	.	.	.	.	3	$\frac{2}{3}$
Type II	.	.	.	.	5	$\frac{1}{2}$
Type III	.	.	.	.	2	$\frac{3}{4}$

Of this set-up we may initially define 3 parameters ( $P_1 = \frac{3}{10}$ ,  $P_2 = \frac{5}{10}$ ,  $P_3 = \frac{2}{10}$ ) respectively specifying the proportionate frequency with which we take a sample at random from any one type of urn; but the frequency with which samples of a particular composition will occur in any one of the three classes of samples so specified will depend on the parameters ( $p_1 = \frac{2}{3}$ ,  $p_2 = \frac{1}{2}$ ,  $p_3 = \frac{3}{4}$ ) which specify the proportion of red balls in the urn definitive of the class. Within the framework of the illustrative (but not necessary) assumption that we sample with replacement, the long run proportionate frequencies with which  $x$  red balls will occur respectively in an  $r$ -fold sample from one or other type of urn are:

$$P_{x.1} = r_{(x)} \cdot \frac{2^x}{3^r}; \quad P_{x.2} = r_{(x)} \cdot \frac{1}{2^r}; \quad P_{x.3} = r_{(x)} \cdot \frac{3^x}{4^r}.$$

Within the framework of the 2-class universe and a 2-stage sampling process, Bayes' Theorem is an exact answer to the question: what is the probability of correctly asserting that an  $r$ -fold sample comes from a sub-universe of type  $M$  if it contains  $x$  items of a class  $A$  (e.g. of red colour in this example)? If we speak of such an assertion as the adoption of hypothesis  $M$ , we may also phrase the issue as: what is the long-run proportionate frequency of correct action based on the assumption that hypothesis  $M$  is true? The answer follows from the product and addition rules, as we shall see below.

It is more easy to appreciate its rationale, if we first set out the result numerically as a balance sheet of long-run frequencies. Let us assume that  $r = 4$  and  $x = 2$ , i.e. that our sample consists of 4 balls of which 2 are red, so that

$$P_{x.1} = \frac{8}{27}; \quad P_{x.2} = \frac{3}{8}; \quad P_{x.3} = \frac{27}{128}.$$

	Proportion of all 4-fold Samples		
	With 2 red balls	Other	Total
From Type I . . .	$\frac{3}{10} \cdot \frac{8}{27} = \frac{12}{135}$	$\frac{3}{10} \cdot \frac{19}{27} = \frac{57}{270}$	$\frac{3}{10}$
From Type II . . .	$\frac{1}{2} \cdot \frac{3}{8} = \frac{3}{16}$	$\frac{1}{2} \cdot \frac{5}{8} = \frac{5}{16}$	$\frac{1}{2}$
From Type III . . .	$\frac{1}{5} \cdot \frac{27}{128} = \frac{27}{640}$	$\frac{1}{5} \cdot \frac{101}{128} = \frac{101}{640}$	$\frac{1}{5}$
Total . . .	$\frac{12}{135} + \frac{3}{16} + \frac{27}{640} = \frac{5505}{17,280}$	$\frac{57}{270} + \frac{5}{16} + \frac{101}{640} = \frac{11,775}{17,280}$	1

If we now abstract from this table the items which refer to samples with the structure specified we derive the following proportion of 4-fold samples with 2 red balls :

Type I	$\frac{12}{135} \div \frac{5505}{17,280} = \frac{512}{1835}$
Type II	$\frac{3}{16} \div \frac{5505}{17,280} = \frac{1080}{1835}$
Type III	$\frac{27}{640} \div \frac{5505}{17,280} = \frac{243}{1835}$
Total	1

We may set out the universe of sampling in more general terms as in Table 1 and the universe of samples as in Table 2. Table 3 then shows the balance sheet of long-run frequencies in the same terms as above. For a formal statement of the theorem we may employ the following symbols :

$p_h = (1 - q_h)$  = proportionate frequency of a successful draw in a unit trial from sub-universe  $H$ .

$P_h$  = proportionate frequency that an  $r$ -fold sample comes from sub-universe  $H$ .

$P_{x.h}$  = conditional proportionate frequency that  $x$  is the score (successes) in the  $r$ -fold sample if it comes from sub-universe  $H$ , so that

$$\text{with replacement :} \quad P_{x.h} = r_{(x)} p_h^x q_h^{r-x};$$

$$\text{without replacement (from sub-universe of } n \text{ items) :} \quad P_{x.h} = r_{(x)} (np)^{(x)} (nq)^{(r-x)} \div n^{(r)}.$$

$P_{hx}$  = proportionate frequency of the combined event that  $x$  is the sample score and that the sample comes from sub-universe  $H$ .

Then by the *product rule*:

$$P_{hx} = P_h \cdot P_{x.h},$$

$P_x$  = proportionate frequency of the event that the score of any  $r$ -fold sample is  $x$ .



Then by the *addition rule*:

$$P_x = \sum_{h=1}^{h=\infty} P_{hx} = \sum_{h=1}^{h=\infty} P_h \cdot P_{x \cdot h},$$

$P_{h \cdot x}$  = proportionate frequency of the conditional event that the  $r$ -fold sample comes from sub-universe  $H$  if  $x$  is the sample score.

Then by the *product rule*:

$$P_{hx} = P_x \cdot P_{h \cdot x}.$$

Whence we obtain the theorem :

$$P_{h \cdot x} = \frac{P_{hx}}{P_x} = \frac{P_h \cdot P_{x \cdot h}}{\sum_{h=1}^{h=\infty} P_h \cdot P_{x \cdot h}} \quad \dots \quad (i)$$

TABLE 1

*A Bayes' Model of a Stratified Universe of Sampling*

Types of Urn	I	II	III	Total
Nos. of each type	$n_1$	$n_2$	$n_3$	$N$
Proportionate contribution <i>ditto to all samples</i>	$P_1 = \frac{n_1}{N}$	$P_2 = \frac{n_2}{N}$	$P_3 = \frac{n_3}{N}$	1
Proportion of <i>red</i> balls in each	$p_1$	$p_2$	$p_3$	...
Frequency of $r$ -fold sample containing $x$ red balls*	$P_{x \cdot 1} = r_{(x)} p_1 (1 - p_1)^{r-x}$	$P_{x \cdot 2} = r_{(x)} p_2 (1 - p_2)^{r-x}$	$P_{x \cdot 3} = r_{(x)} p_3 (1 - p_3)^{r-x}$	$P_x$

TABLE 2

*The Bayes' Model for  $r$ -fold samples*

From Urn	With $x$ red balls	With more or less than $x$ red balls	Total
I	$P_1 \cdot P_{x \cdot 1}$	$P_1 (1 - P_{x \cdot 1})$	$P_1$
II	$P_2 \cdot P_{x \cdot 2}$	$P_2 (1 - P_{x \cdot 2})$	$P_2$
III	$P_3 \cdot P_{x \cdot 3}$	$P_3 (1 - P_{x \cdot 3})$	$P_3$
Total	$P_x = \sum_{h=1}^{h=3} P_h \cdot P_{x \cdot h}$	$1 - P_x$	1

\* If sampling is without replacement we write for an urn containing  $u_m$  balls of which  $s_m$  are red

$$L_m = r_{(x)} s_m^{(x)} (u_m - s_m)^{(r-x)} \div u_m^{(r)}.$$

TABLE 3

*Proportionate Frequency of the event that the r-fold sample contains x red balls*

Urn	Comes from	Does <i>not</i> come from	Total
I	$\frac{P_1 \cdot P_{x \cdot 1}}{P_x} = P_{1 \cdot x}$	$1 - P_{1 \cdot x}$	1
II	$\frac{P_2 \cdot P_{x \cdot 2}}{P_x} = P_{2 \cdot x}$	$1 - P_{2 \cdot x}$	1
III	$\frac{P_3 \cdot P_{x \cdot 3}}{P_x} = P_{3 \cdot x}$	$1 - P_{3 \cdot x}$	1
Total	1	1	1

In Table 3  $P_{h \cdot x}$  is the proportionate frequency of the event that an  $r$ -fold sample with score  $x$  comes from an urn of type  $H$ , i.e. the long-run frequency, among all samples so constituted, of those which come from such an urn. To assert that an  $r$ -fold sample does so on the basis of the additional information that the sample score is  $x$ , is to assert that hypothesis  $H$  is true; and  $P_{h \cdot x}$  is the proportion of such assertions which correctly describe what happens in the long run. Thus we may re-interpret (Table 4) the items of Table 3 as probabilities of the truth or falsehood of the assertion that a particular hypothesis is correct. Any items of the form  $(1 - P_{h \cdot x})$  in the column headed *False* thus correspond to what we have called in 20.00 the *uncertainty safeguard of the unconditional assertion that hypothesis H is true*.

TABLE 4

*Long-run frequency of Statements about the r-fold sample containing x red balls*

From Urn	True	False	Total
I	$P_{1 \cdot x} = \frac{P_1 \cdot P_{x \cdot 1}}{P_x}$	$1 - P_{1 \cdot x}$	1
II	$P_{2 \cdot x} = \frac{P_2 \cdot P_{x \cdot 2}}{P_x}$	$1 - P_{2 \cdot x}$	1
III	$P_{3 \cdot x} = \frac{P_3 \cdot P_{x \cdot 3}}{P_x}$	$1 - P_{3 \cdot x}$	1

Before proceeding further it may be helpful to some readers if we first pause to dispose of a common difficulty. The logic of Bayes' theorem is not obscure or subtle against the background of an urn model; but appreciation of the relevance of the model to statistical inference in the domain of the world's work makes no mean demands upon the imagination.\* A biological illustration may assist the reader who boggles at this step. We shall suppose that a culture of fruit flies contains 100 females of which 5 carry a sex-linked lethal gene. Concerning one of these flies we know that  $\frac{2}{5}$  of its progeny are female. Now this will occasionally happen if it is normal, but much less rarely if it carries a sex-linked lethal gene. If we merely know

\* To add to the difficulties of the plain man, current and authoritative works repeat such paradoxical definitions as that the prior probability of the hypothesis is the probability assigned thereby to the "event before it has happened". Actually, our prior and posterior probabilities refer to different events, one to the probability that *any*  $r$ -fold sample comes from a sub-universe of type  $H$  and one to the probability that a particular sub-class of such samples (i.e. those with score  $x$  in our model set-up) comes from it.



that an individual is a female we know that there is a 95 per cent. chance that she is normal. This is the *prior probability of the null hypothesis* that the female is normal. The null hypothesis assigns as the probability that any single offspring of a normal fly will be female  $p_a = \frac{1}{2}$ . The alternative hypothesis that she carries a sex-linked lethal gene assigns as the probability that any single offspring will be female  $p_b = \frac{2}{3}$ . If our illustrative mother fly has 100 offspring of which three-fifths (i.e. 60) are female, the frequencies of correct judgments based on the assumption of normality or otherwise will be in the ratio

$$\frac{95}{100} \cdot \frac{100!}{60! 40!} 2^{-100} : \frac{5}{100} \cdot \frac{100!}{60! 40!} 3^{100}.$$

This is approximately 6.75 : 1 in favour of the null hypothesis. If we estimate the relative frequencies of correct decisions on the false assumption that there are just as many female flies of both sorts in the culture, we should obtain the figure 0.36 : 1 or about 3 : 1 against the null hypothesis. Such an assumption, known as Bayes' postulate, is on all fours with a common misconception implicit in the lay-out of age-case distributions in extant medical textbooks. Peptic ulcer is a much more common complaint after 40 than before that age ; but it would be a fallacy to assert that a conscript is over 40 because he has peptic ulcer. The proportion of conscripts over forty is very much less than that of younger men. Hence the actual number of younger men with peptic ulcer may well be greater than the actual number of men over 40 years of age. If so, there are more conscripts with peptic ulcer of whom one can correctly assert that they are not yet 40 years old.

In short, Bayes' postulate is the *vulgar error of neglecting the population at risk*. The true prior probabilities which it gratuitously equalises are in this context the age-standardising weights which make the balance sheet of risk a true bill. Though we may not be able to attach an exact figure to them, we may be able to set some agreed limit on their relative values. In any case, the circumstance that we cannot do so with assurance constitutes no justification for assuming equality. The fact is that one undertakes an experiment to test a hypothesis either because one has good reason to believe in its truth or because one has good reason to suspect its falsity.\* Good investigators do not commonly undertake experiments unless they have one or other end in view, i.e. unless there is factual basis for the belief that the prior probabilities are unequal.

Much misunderstanding arises through speaking of the prior probability of a hypothesis when we cannot indeed distinguish between a hypothesis which specifies a parameter definitive of an *existent* population at risk and a hypothesis which specifies a parameter definitive of one which conceivably might exist. To be sure we can say that its prior probability is zero if the population specified by it is non-existent ; but the distinction, if formally trivial in this sense, is useful in another. Sometimes, as in the Model I situation of 20.06 below, we may postulate a sampling process which involves only one level of choice. By definition, we may then assign unity as prior probability to the correct hypothesis and zero as that of any conceptual alternative. In the general model situation—Model II of 20.08—to which Bayes' balance sheet is relevant we conceive a 2-stage sampling process, the first being the choice of the urn or of the individual fruitfly in our previous examples. The impossibility of assigning in most real situations appropriate numerical values to the prior probabilities is thus only one horn of the dilemma with

\* F. J. Anscombe (1951), *Mind*, Vol. 60, rightly comments as follows :

"As soon as any proposition or hypothesis has been formulated which is worth testing experimentally, there is already evidence as to its truth derived from existing accepted knowledge and from considerations of analogy or 'consilience'. A question to which we have no grounds whatever for hazarding an answer is an idle question and would not be the subject of scientific investigation."







postulate : (i) two types of urn,  $A$  containing red and black balls in the ratio 1 : 1 and  $B$  containing red and black balls in the ratio 2 : 1 ; (ii) sampling with replacement, the score  $x$  being the number of red balls in the  $r$ -fold sample. Thus our hypotheses are

$$\text{Hypothesis } A \quad p_a = \frac{1}{2}.$$

$$\text{Hypothesis } B \quad p_b = \frac{2}{3}.$$

Whence we have

$$S_{ab \cdot r} = \frac{\left(\frac{1}{2}\right)^x \left(\frac{1}{2}\right)^{r-x}}{\left(\frac{2}{3}\right)^x \left(\frac{1}{3}\right)^{r-x}} = \frac{3^r}{2^{r+x}} \quad \dots \dots \dots (iv)$$

From (ii) we can proceed to tabulate results of score sequences involving different total scores in an 8-fold sample as below :

Score ( $x$ )	Sequential Ratio ( $S$ )	Score ( $x$ )	Sequential Ratio ( $S$ )
0	$\frac{6 \ 5 \ 6 \ 1}{2 \ 5 \ 6} = 25.6$	5	$\frac{6 \ 5 \ 6 \ 1}{8 \ 1 \ 9 \ 2} = 0.801$
1	$\frac{6 \ 5 \ 6 \ 1}{5 \ 1 \ 2} = 12.8$	6	$\frac{6 \ 5 \ 6 \ 1}{1 \ 6 \ 3 \ 8 \ 4} = 0.400$
2	$\frac{6 \ 5 \ 6 \ 1}{1 \ 0 \ 2 \ 4} = 6.41$	7	$\frac{6 \ 5 \ 6 \ 1}{3 \ 2 \ 7 \ 6 \ 8} = 0.200$
3	$\frac{6 \ 5 \ 6 \ 1}{2 \ 0 \ 4 \ 8} = 3.20$	8	$\frac{6 \ 5 \ 6 \ 1}{6 \ 5 \ 5 \ 3 \ 6} = 0.100$
4	$\frac{6 \ 5 \ 6 \ 1}{4 \ 0 \ 9 \ 6} = 1.60$		$\dots \dots \dots$

From inspection of the above we see that a critical score level  $x_c = 4.5$  divides all 8-fold samples into two sets. If  $x < x_c$ ,  $S_{ab \cdot 8} > 1$ , i.e. hypothesis  $A$  assigns a higher probability to the observed sequence than does hypothesis  $B$ . If  $x > x_c$ ,  $S_{ab \cdot 8} < 1$  and the converse is true. Without making any claims which sidestep the Bayes' dilemma, we may choose to be content with : (i) accepting hypothesis  $A$  if it assigns a probability nine times as great as does hypothesis  $B$  to the observed sequence, i.e.  $S_{ab \cdot 8} \geq 9$  ; (ii) accepting hypothesis  $B$  if the converse is true, i.e.  $S_{ab \cdot 8} \leq 0.1$ . If so, we may define 3 score levels as follows :

$$\begin{aligned} x &< 1.5 && \text{content to accept } A. \\ 1.5 &< x < 7.5 && \text{no decision} \\ x &> 7.5 && \text{content to accept } B. \end{aligned}$$

Let us now suppose that our  $x$  score is in fact 2 at the 8th trial. If so, we suspend judgment. At the next (9th) trial the total score must be 2 or 3 ; and

$$\begin{aligned} S_{ab \cdot 9} &= \frac{3}{2} S_8 \quad \text{or} \quad \frac{3}{4} S_8, \\ \therefore S_{ab \cdot 9} &= \frac{19683}{2048} \quad \text{or} \quad \frac{19683}{4096}. \end{aligned}$$

On the assumption that we accept hypothesis  $A$  when  $S_{ab \cdot 9} \geq 9$  we shall thus reach a decision at the 9th trial if we then score a failure (extract a black ball) and suspend judgment if we then score a success. At each trial we may in fact suspend judgment without bias to subsequent decision.

If we now return to (iii) we can give a meaning to our test criterion in terms of Bayes' prior probabilities. To say that we shall be more often right than wrong if we act on the assumption that hypothesis  $A$  is true than if we act on the assumption that the alternative hypothesis is true is equivalent to writing  $B_{ab \cdot r} > 1$ , whence

$$k_{ab} > \frac{1}{S_{ab \cdot r}} \quad \text{or} \quad \frac{P_a}{P_b} > \frac{1}{S_{ab \cdot r}} \quad \dots \dots \dots (v)$$

If we make our criterion of acceptance of hypothesis  $A$  that  $S_{ab..r} \geq 9$  as above, we therefore mean that we should be right in accepting it more often than in accepting hypothesis  $B$  so long as the prior probability of the latter does not exceed 9 times that of the former. If we make our criterion for rejecting hypothesis  $A$  (accepting the alternative) that  $S_{ab..r} \geq 0.1$  we signify that we shall be more often right than wrong in doing so unless hypothesis  $A$  has a prior probability more than 9 times that of hypothesis  $B$ .

The use of the sequential ratio so prescribed is different from other test procedures because it permits us to proceed to a decisive and unbiased verdict without prescribing in advance how large a sample will be necessary. This, of course, presupposes a positive answer to the question: can we guarantee that the outcome will eventually be decisive if we make the sample size ( $r$ ) sufficiently large?

Before we attempt to answer this question, it will be instructive to formulate more precisely a conclusion already stated: if  $S_r$  and  $S_{r+1}$  are sequential ratios for the  $r$ th and the  $(r+1)$ th sample respectively, within what limits does  $S_{r+1}$  lie? Let us write  $p_b = mp_a$ , whence

$$S_r = \frac{p_a^x (1 - p_a)^{r-x}}{m^x \cdot p_a^x (1 - mp_a)^{r-x}},$$

$$\therefore S_r = \frac{1}{m^x} \cdot \left( \frac{1 - p_a}{1 - mp_a} \right)^{r-x} \quad \text{. . . . . (vi)}$$

Only two cases may arise when we enlarge the sample from  $r$  to  $r+1$  items: (i)  $x$  may remain fixed if the result of the further trial is a failure; (ii)  $x$  may increase by unity if the result is a success. Thus we may put

$$S_{r+1} = \frac{1}{m^x} \left( \frac{1 - p_a}{1 - mp_a} \right)^{r-x+1} \quad \text{or} \quad \frac{1}{m^{x+1}} \left( \frac{1 - p_a}{1 - mp_a} \right)^{r-x} \quad \text{. . . . . (vii)}$$

Thus we have

$$\frac{S_{r+1}}{S_r} = \frac{1 - p_a}{1 - mp_a} \quad \text{or} \quad \frac{1}{m} \quad \text{. . . . . (viii)}$$

In particular, when  $p_a = \frac{1}{2}$  we have

$$S_r = \frac{1}{m^x (2 - m)^{r-x}};$$

$$S_{r+1} = \frac{1}{m^x (2 - m)^{r-x+1}} \quad \text{or} \quad \frac{1}{m^{x+1} (2 - m)^{r-x}},$$

$$\therefore \frac{S_{r+1}}{S_r} = \frac{1}{2 - m} \quad \text{or} \quad \frac{1}{m} \quad \text{. . . . . (ix)}$$

Since  $p_b \leq 1$ ,  $m \leq p_a^{-1}$  in (v) and  $m \leq 2$  in (vi). When  $p_b = \frac{2}{3}$  and  $p_a = \frac{1}{2}$ ,  $m = \frac{4}{3}$ ; and (ix) becomes

$$\frac{S_{r+1}}{S_r} = \frac{3}{2} \quad \text{or} \quad \frac{3}{4} \quad \text{. . . . . (x)}$$

As stated, a sequential test would be of little value if we had no assurance that it would eventually terminate, i.e. that the sequential ratio will attain an upper limit  $A$  assigned as our criterion that hypothesis  $A$  is acceptable and a lower limit  $B$  assigned as our criterion that



hypothesis  $B$  is acceptable. For simplicity we may adopt the convention that  $m > 1$  in (vi) *et seq.*, i.e.  $p_a < p_b$ , and examine the consequences of the empirical rule of thumb commonly called the law of the constancy of great numbers, *viz.* that  $x$  approaches its expected mean value  $rp$  more closely as we make  $r$  larger. Thus we may say:

- (i) if hypothesis  $A$  is true,  $x$  eventually approaches indefinitely near the limit  $x = rp_a$ ;
- (ii) if hypothesis  $B$  is true,  $x$  eventually approaches indefinitely near the limit

$$x = rp_b = mrp_a.$$

Let us therefore examine what values  $S_a$ ,  $S_b$ , respectively the sequential ratio takes when hypothesis  $A$  and hypothesis  $B$  are true, and  $r$  is very large in this sense.

For the arithmetical example already cited  $p_a = \frac{1}{2}$ , and  $m = \frac{4}{3}$ , so that

$$S = \left( \frac{2-m}{m} \right)^x \cdot \frac{1}{(2-m)^r} = \frac{3^r}{2^{x+r}}.$$

The two limiting values of  $x$  are  $\frac{1}{2}r$  and  $\frac{2}{3}r$ , so that

$$S_a = \frac{3^r}{2^{\frac{3r}{2}}} = \left( \frac{9}{8} \right)^{\frac{r}{2}} \quad \text{and} \quad S_b = \frac{3^r}{2^{\frac{5r}{3}}} = \left( \frac{27}{32} \right)^{\frac{r}{3}}.$$

When  $r$  is indefinitely large  $S_a$  itself approaches infinity and  $S_b$  approaches zero. It will suffice to formalise this for a particular case of special interest, *viz.*  $p_a = \frac{1}{2}$ ,  $1 > p_b > p_a$ , so that  $1 < m < 2$ . We then have

$$S_a \simeq \left( \frac{2-m}{m} \right)^{\frac{1}{2}r} \cdot \frac{1}{(2-m)^r} = \frac{1}{(2m-m^2)^{\frac{1}{2}r}} \quad . \quad . \quad . \quad (xi)$$

$$S_b \simeq \left( \frac{2-m}{m} \right)^{\frac{rm}{2}} \cdot \frac{1}{(2-m)^r} = \frac{(2-m)^{\frac{1}{2}(mr-2r)}}{m^{\frac{1}{2}mr}} \quad . \quad . \quad . \quad (xii)$$

Since  $m$  lies inside the limits 1 and 2,  $(2m - m^2)$  in (xi) lies inside the limits 0 and 1, so that  $S_a$  in (xi) is indefinitely large when  $r$  is also. The meaning of (xii) is more clear, if we write it as

$$S_b = \frac{1}{m^r} \left( \frac{2-m}{m} \right)^{\frac{1}{2}(mr-2r)}.$$

Since  $m > 1$ , the first factor in the above becomes smaller and smaller when  $r$  becomes larger. That this is also true of the second factor is evident, if we write  $m = 1 + h$  in which  $h$  is positive and less than unity if  $p_b < 1$ . Thus we have

$$\frac{2-m}{m} = \frac{1-h}{1+h} < 1.$$

Evidently therefore  $S_b$  approaches zero as its limiting value.

If then the hypotheses  $p_a = \frac{1}{2}$  and  $p_b = \frac{1}{2}m$  constitute an *exclusive and exhaustive* set in the sense that one is true if the other is false and *vice versa*, the foregoing reasoning leads to the conclusion that the test will terminate in the rejection of the false and the acceptance of the true one if we make  $r$  sufficiently large. This does *not* imply that it will do so if a third hypothesis is admissible.

A single example will suffice to make this clear. Let us again suppose that we set out to test the alternative hypotheses  $p_a = \frac{1}{2}$  and  $p_b = \frac{2}{3}$ . This time, however, we shall admit the possibility that  $p_c = \frac{3}{5}$  is the true value of the sub-universe parameter  $p$ , so that  $x$  tends to  $\frac{3}{5}r$  in the limit. Thus the limiting value of the sequential ratio will be

$$\frac{3^r}{2^{\frac{8r}{5}}} = \left( \frac{243}{256} \right)^{\frac{r}{5}}.$$

This ratio becomes smaller and smaller as  $r$  becomes larger. Hence the test will eventually lead wrongly to acceptance of the hypothesis  $p_b = \frac{2}{3}$  in preference to  $p_a = \frac{1}{2}$ . Alternatively, we may admit the possibility  $p_a = \frac{1}{2}$  as the true value of  $p$ . Whence we get the limiting sequential ratio as

$$\left( \frac{3^{20}}{2^{31}} \right)^{\frac{r}{20}} \simeq (1.63)^{\frac{r}{20}}.$$

This ratio becomes larger as  $r$  increases. Hence the test will eventually lead wrongly to the acceptance of the hypothesis  $p_a = \frac{1}{2}$  in preference to  $p_b = \frac{2}{3}$  if the true value of  $p$  is 0.55.

The exact boundary is definable in terms which the reader can generalise. For the particular case when our sequential ratio is referable to the hypotheses  $p_a = \frac{1}{2}$  and  $p_b = \frac{2}{3}$ , we shall postulate a true value  $p_c = \frac{1}{2}k$  so that  $x$  tends to the limit  $\frac{1}{2}kr$  and  $S$  to the limit

$$\frac{3^r}{2^{\frac{1}{2}r(k+2)}} = S_k = \left( \frac{9}{2^{k+2}} \right)^{\frac{r}{2}}.$$

The test will not terminate if  $S_k = 1$ , i.e. if  $2^{k+2} = 9$ , and

$$\log 9 = (k+2) \log 2,$$

$$\therefore k = \frac{\log 9}{\log 2} - 2 \simeq \frac{117}{100}.$$

If we set  $p_a = \frac{1}{2}$  and  $p_b = \frac{2}{3}$ , the sequential ratio will diminish as  $r$  increases if  $p > \frac{1}{2}\frac{17}{10}$  and increases as  $r$  increases if  $p < \frac{1}{2}\frac{17}{10}$  as the two examples last cited show. The reader should not find it difficult to generalise this result, the implications of which are clear. The interpretation of the test procedure against the background of the Bayes' ratio presupposes that each alternative hypothesis has a finite prior probability. If both hypotheses are wrong and both are inadmissible, the test will terminate in a wrong decision. This raises the question: can we formulate the sequential test procedure in terms of alternative hypotheses which cannot both be wrong?

We have already seen that we can interpret an inspection plan in terms of the alternative hypotheses  $p \leq p_a$  and  $p \geq p_b$ . These do not constitute an exclusive set, since it is possible that  $p_a < p < p_b$ ; but we can agree to confine the verdict of the test procedure to the denial of one or the other, i.e. to alternative statements to the effect  $p > p_a$  or  $p < p_b$ . Neither statement is then inconsistent with the possibility last stated. Thus a test procedure will result in denying one of the propositions  $p = \frac{1}{2}$  and  $p = \frac{2}{3}$  if designed to ensure the negation of one or other of the hypotheses  $p \leq \frac{1}{2}$  and  $p \geq \frac{2}{3}$ ; but we have not as yet shown how it is possible to interpret sequential ratio limits unless the alternatives assume the exact form  $p = p_a$  or  $p = p_b$ . This will be the theme of 20.05 below. The advantage of doing so, if possible, is that we can continue to sample until we have reached a decision without prescribing in advance what size of sample will necessarily ensure a conclusive outcome.



## 20.03 LIMITATIONS OF THE UNIQUE NULL HYPOTHESIS

For the past generation research workers engaged on agricultural trials, tests of the efficacy of therapeutic or prophylactic measures, sociological field work and bioassay have relied for validification of their results on decision tests devised by R. A. Fisher and his co-workers in conformity with a familiar pattern. Such tests entail: (a) the invocation of a unique so-called *null* hypothesis which prescribes the frequency with which a sample score will lie outside a prescribed limit (or limits); (b) the specification of a criterion of rejection, i.e. the convention to reject the hypothesis if the sample score does in fact lie outside the prescribed limit(s). Customarily (and oddly) the corresponding limiting frequency adopted is at the 95 per cent. (approximately 20 : 1 odds) level; and the possibility of defining it in such terms resides in the fact that the unique hypothesis chosen for the purpose has an assignable distribution function. With one notable exception the latter is specifiable, though rarely recognised as such, within the framework of the Type System of Karl Pearson.\*

From one viewpoint, the prevalence of the fashion referred to is comprehensible. The publication of *Statistical Methods for Research Workers* prepared the way for manuals by Snedecor, Tippett, Hagood, Quenouille and others, exhibiting schemata for computation in conformity with the Fisher test prescriptions. By recourse to a wealth of exemplary material the research worker willing to take the test prescription on trust can therefore readily, it may be all too readily, select a type specimen at least seemingly like his or her own problem. None the less, there must be among those who do so, not a few who have felt misgiving for any (or all) of several reasons, notably the following:

- (a) not infrequently the form of the null hypothesis is irrelevant to the main issue, e.g. as when the decision that two treatment procedures have different results is of trivial interest in comparison with the decision that treatment *B* is at least *so much* more effective than treatment *A*;
- (b) the type of decision which concerns the investigator determines the choice of a particular null hypothesis far less than considerations of algebraic convenience *vis-à-vis* the specification of a sample distribution;
- (c) the test prescriptions take no stock of any alternative hypothesis which may indeed be the main concern of the investigator.

The first misgiving has special reference to the domain of estimation, and as such to the theory of confidence specially associated with the names of J. Neyman and E. S. Pearson. The second and third raise issues which a theory of test procedure also advanced by Neyman and Pearson has brought into focus; but their critique of the unique null hypothesis has to date exerted little influence on research workers outside America. This is less because their writings lack the polemic vitality of their predecessors than because the concepts invoked are logically subtle and on that account difficult to assimilate unless examined against a background of familiar material. The aim of what follows is to help the laboratory or the field research worker to recognise pitfalls in previously accepted test procedures and to materialise some of the essentially novel concepts of the Neyman-Pearson approach.

\* Thus Snedecor's variance ratio *F*-test described as a score transformation of Fisher's *z*-test is really a type VI, the Gosset *t*-test a type VII, the distribution of the sample variance a type III, including as a special case the current test for the  $2 \times 2$  Table (1 d.f.) which is (as Fisher himself first pointed out) formally equivalent to the normal proportionate score difference test in the binomial domain, the significance test for the correlation ratio and for Spearman's rank correlation coefficient is a type II, and the best fitting curve for a non-replacement sampling distribution in the 2-class universe is a type I. Fisher's distribution of the product-moment index for non-zero covariance when regression is linear is the notable exception to the foregoing remarks.



One way to do so is to formulate a biological problem which involves no predilection for a single hypothesis as the *null* one in virtue of algebraic convenience as such. Accordingly, we shall again (p. 852) think of a culture of *Drosophila* containing normal females and females which carry a sex-linked lethal gene. With Bacon we shall concede that nature is more diverse in her operations than man in his conceptions, but our knowledge of laboratory conditions (presumptively highly standardised) will justify the provisional assumption that any such female fruit-fly with an excessively large number of female offspring will, in fact, be either an entirely normal female or a lethal carrier. That is to say, we exclude such a contingency as the possibility that there is an endemic rare virus disease more fatal to male than to female larvae. We may then with some justifiable assurance postulate two hypotheses about any female in the culture:

Hypothesis *A*: the female is normal, in which event the probability that any one of her offspring will be female is  $p_a = \frac{1}{2}$ .

Hypothesis *B*: the female is a lethal carrier, whence the probability that any one of her offspring is female is  $p_b = \frac{2}{3}$ .

We shall now suppose that a particular female has 144 offspring, and examine the current theory of test procedure when the end in view is to decide whether we shall adopt one or other hypothesis. Our primary concern will thus be with what the test prescribes, and as such has no necessary connexion with whether it leads us to a correct decision.

We first note that each hypothesis equally prescribes for 144-fold fraternities referable to a single fly mother the long-run frequency of such as respectively contain 0, 1, 2 . . . 143, 144 females. We may specify the relevant parameters thus

	Size of sample (fraternity)	Probability that any single offspring is female	Mean no. of females in sample fraternity	s.d. of score distribution of the sample
<i>A</i>	144	$p_a = \frac{1}{2}$	$M_a = 72$	$\sigma_a = 6$
<i>B</i>	144	$p_b = \frac{2}{3}$	$M_b = 96$	$\sigma_b = 5.66$

From an algebraic viewpoint neither hypothesis specified above has anything to commend it as preferable to the alternative; but we may lazily and arbitrarily agree to consider first of all the consequences of adopting *A* as the *null hypothesis* in the traditional sense, if only because laboratory and field workers would commonly do so in a comparable situation. Lazily and arbitrarily also, we shall first adopt a *modular* criterion of rejection for the same reason, i.e. we shall reject the hypothesis chosen unless the number of females  $x$  is such that

$$| (x - M_a) | \leq X_a.$$

In conformity likewise with current convention, we shall choose the score  $X_a$  so that the probability ( $\alpha$ ) that  $x$  will lie in the *critical region*, i.e. outside the range specified above is about 0.05 if the *null hypothesis correctly describes the situation*. For samples of 144 and values of  $p_a$  (or  $p_b$ ) anywhere (as in this example) within the range 0.1 to 0.9, the normal integral gives an adequate quadrature at the so-called 96 per cent. significance level, if we make the appropriate half interval correction. If we choose  $X_a = +12.5$ , so that  $(x - M_a) = \pm X_a$  when  $(x - M_a) \simeq \pm 2.08\sigma$  the table of the normal integral sets  $\alpha \simeq 0.038$ . In effect, we now have made the decision to regard the female with 144 offspring as normal if the number of her female offspring lies in the range 60 to 84 inclusive and to reject her claims as such, i.e. in this context to regard her as a lethal carrier, if her female offspring number more than 84 or less than 60.

In the last sentence we assume that the end in view of the test is to arrive at a decision, as do at least ninety-nine per cent. of laboratory workers who invoke it in communicating the results



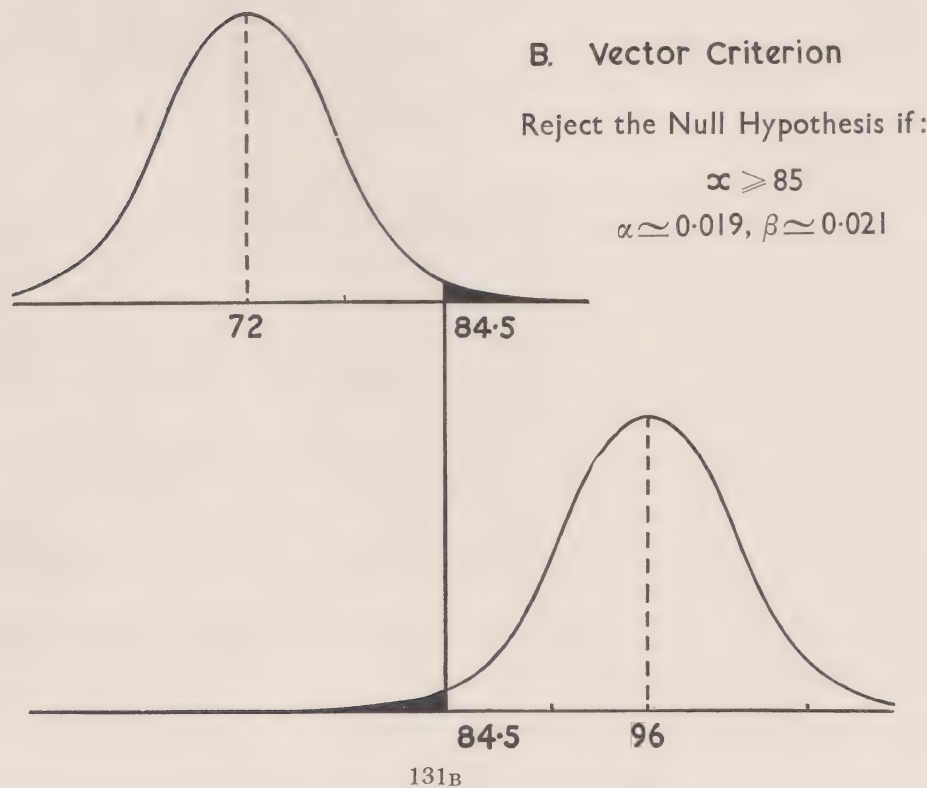
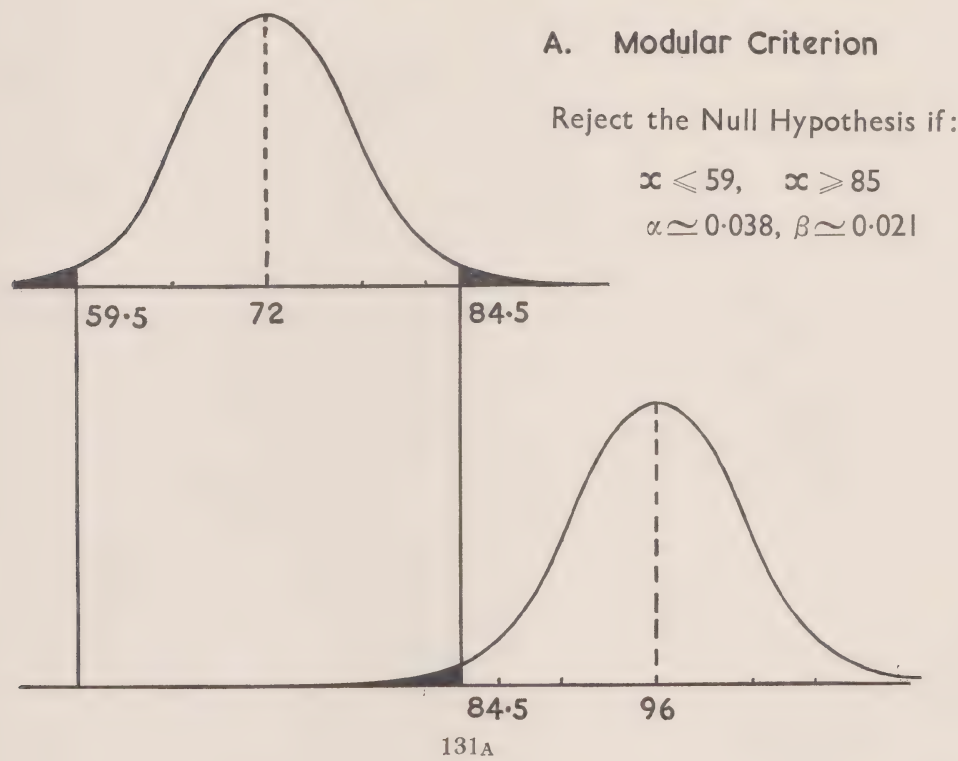


FIG. 131. Testing a null hypothesis ( $p_a = 0.5$ ) against the background of a single admissible alternative hypothesis B ( $p_b = 0.6$ ).

of their researches to the world at large. Fisher himself (*vide Design of Experiments*, fifth edition, 1949, p. 16) says: 'It should be noted that the null hypothesis is never proved or established but is possibly disproved in the course of experimentation.' A doctrinaire exponent of the Yule-Fisher test procedure is therefore free to disclaim the intention stated in favour of an affirmative or positive answer in contradistinction to the alternative verdicts: (a) hypothesis false; (b) hypothesis unproven. If so, only one of two sorts of error we shall now distinguish is relevant to the outcome of the test; but the evasion of the other forces us to a somewhat damaging admission mentioned at the end of 20.04 and more fully in 20.08.

In the Neyman-Pearson theory of test procedure we speak of  $\alpha$  as the conditional probability of making an *error of the first kind*. Now the cited value of  $\alpha$  ( $\simeq 0.038$ ) correctly assigns the probability of rejecting the null hypothesis only on the assumption that the latter is true, i.e. that the mother fly is normal. This we do not know, the aim of the test being to throw light on the alternative possibility. If we carry out the rule of the test consistently, we shall sometimes make an error of the first kind, i.e. reject normal flies as such and by the same token wrongly identify as carriers flies which are indeed normal. Conversely, we shall sometimes apply the test to flies which are indeed carriers. If the number of females among their progeny lies within the range 60–84 inclusive we shall reject them as such. We shall then wrongly accept the null hypothesis. This is the *error of the second kind*, which we make in this context if the relevant parameters of the appropriate distribution are  $p_b = \frac{2}{3}$ ,  $M_b = 96$  and  $\sigma_b \simeq 5.66$ . With due regard to the half interval correction, the region we then exclude is from 59.5 to 84.5 bounded by  $(x - M_b) = -36.5$  and  $(x - M_b) = -11.5$ , i.e.  $(x - M_b) \simeq -6.0\sigma_b$  and  $(x - M_b) \simeq -2.03\sigma_b$ . Since the area of the normal integral of unit variance from  $-\infty$  up to  $-6.4$  is utterly trivial, we make no sensible error if we say that the consistent application of the rule leads us now to reject carriers as such with a probability ( $\beta$ ) assigned by the area of the normal curve of unit variance in the range from  $-\infty$  to  $-2.03$ . We speak of this loosely as the probability of making an error of the second kind; and the table of the normal integral in this case cites the value  $\beta \simeq 0.0212$ . More explicitly,  $\beta$  is the *conditional* probability of accepting the null hypothesis, *if* it is false; but we have as yet said nothing about how often it will be false.

In short, the only information we have at our disposal so far bears on the probability ( $\alpha$ ) of rejecting the null hypothesis when it is true, and that ( $\beta$ ) of rejecting the alternative when the latter is true, i.e. of accepting the null hypothesis *when* it is false. If we now suppose that we actually know the proportion of normal and carrier females in the culture, we can take our analysis a decisive step forward. We shall assume that the culture consists of 500 mothers of which 450 are normal and 50 are carriers. If we then choose at random \* any single fly with 144 offspring as a test subject we may say that

- (i)  $P_a = 0.9$  is the probability that it will be normal, i.e. the probability that the null hypothesis is applicable to the test subject;
- (ii)  $P_b = 0.1 = (1 - P_a)$  is the probability that it will be a carrier, i.e. that the alternative hypothesis is applicable to the test subject.

We now have all the relevant data for a statistical specification of the long-run frequency of all 4 possible results of the outcome of the test:

The fly is normal and we rightly accept it as such

$$P_a(1 - \alpha) = (0.9)(0.962) = 0.866.$$

\* The effect of the lethal gene on the fertility of the fly introduces a bias for which we can allow, and one which we may therefore deliberately neglect for *heuristic* purposes.



$$P_a \cdot \alpha = (0.9) (0.038) = 0.034.$$
$$(1 - P_a)(1 - \beta) = (0.1)(0.979) = 0.098.$$
$$(1 - P_a)\beta = (0.1)(0.021) = 0.002.$$

	Assertion true	Assertion false
Null hypothesis true	$P_a(1 - \alpha) = 0.866$	$P_a \cdot \alpha = 0.034$
Null hypothesis false	$(1 - P_a)(1 - \beta) = 0.098$	$(1 - P_a)\beta = 0.002$
Total	$1 - \beta - (\alpha - \beta)P_a = 0.964$	$\beta + (\alpha - \beta)P_a = 0.036$

$$P_t = 1 - \beta - (\alpha - \beta)P_a = 96.4 \text{ per cent.} \quad (i)$$

[illegible]

$$P_t = 0.981 = 98.1 \text{ per cent.}; P_f = 0.019 = 1.9 \text{ per cent.}$$

At this stage we may also note with profit an interesting consequence of (i). If  $\alpha = \beta$  so that  $1 - \beta = 1 - \alpha$  and  $(\alpha - \beta) = 0$ , equation (i) reduces to  $P_t = (1 - \alpha)$ . Within the framework of our assumptions that there is only one admissible alternative to the null hypothesis,

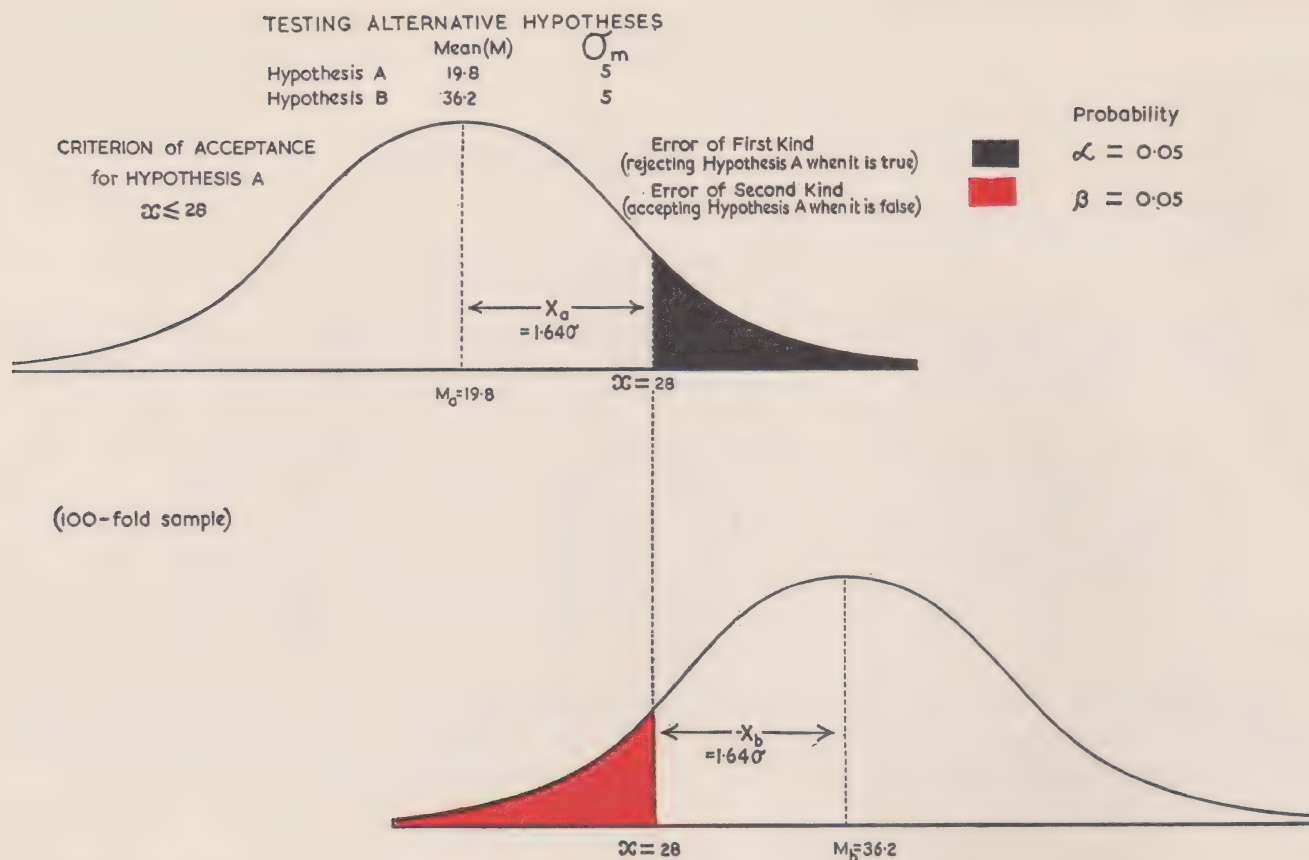


FIG. 132. Testing exclusive Alternative Hypotheses : (i) Rejection-Acceptance Criterion chosen to make error of first kind equal to error of second kind.

so that  $P_b = (1 - P_a)$ , we can assign a value to the long run frequency of correct decision based on consistent application of the rule without any prior knowledge ( $P_a$  or  $P_b$ ) of the population at risk if we *define our rejection criterion in such a way as to equalise the probabilities of errors of the two kinds*. We can then predetermine that the value of  $\alpha$  may be as small as we care to make it by prescribing a *sample size sufficiently large*. Needless to say, this presupposes the possibility of defining the distribution function of the single admissible alternative hypothesis.

Within the framework of assumptions and in the same model set-up, let us now explore the effects of making our criterion for rejecting the null hypothesis more exacting in the sense that our error of the first kind is less. Thus we shall decide to accept a female fly with 144 offspring as normal if (vector criterion) she has 88 or less female offspring and deem her (rightly or wrongly) to be a lethal carrier if she has 89 or more. We then set the decision limits on either side of  $x = 88.5$ , in which event  $(x - M_a) = +2.75\sigma_a$  and  $(x - M_b) = -1.33\sigma_b$ . Whence from the table of the normal integral we derive  $\alpha = 0.003$  and  $\beta = 0.092$ . If we paint these values in (i) we get

$$1 - \beta = 0.908 \quad \text{and} \quad (\alpha - \beta)P_a = -0.080,$$

$$\therefore P_t = 0.988 \quad \text{or} \quad 98.8 \text{ per cent.}$$

In this situation little advantage (98.8 as against 98.1 per cent.) accrues from making our criterion for rejection of the null hypothesis more exacting; but we have chosen our null hypothesis as the hypothesis with greater *prior probability* since the culture contains a great excess of normal flies. Let us then reverse the situation by postulating that the culture contains 450 lethal



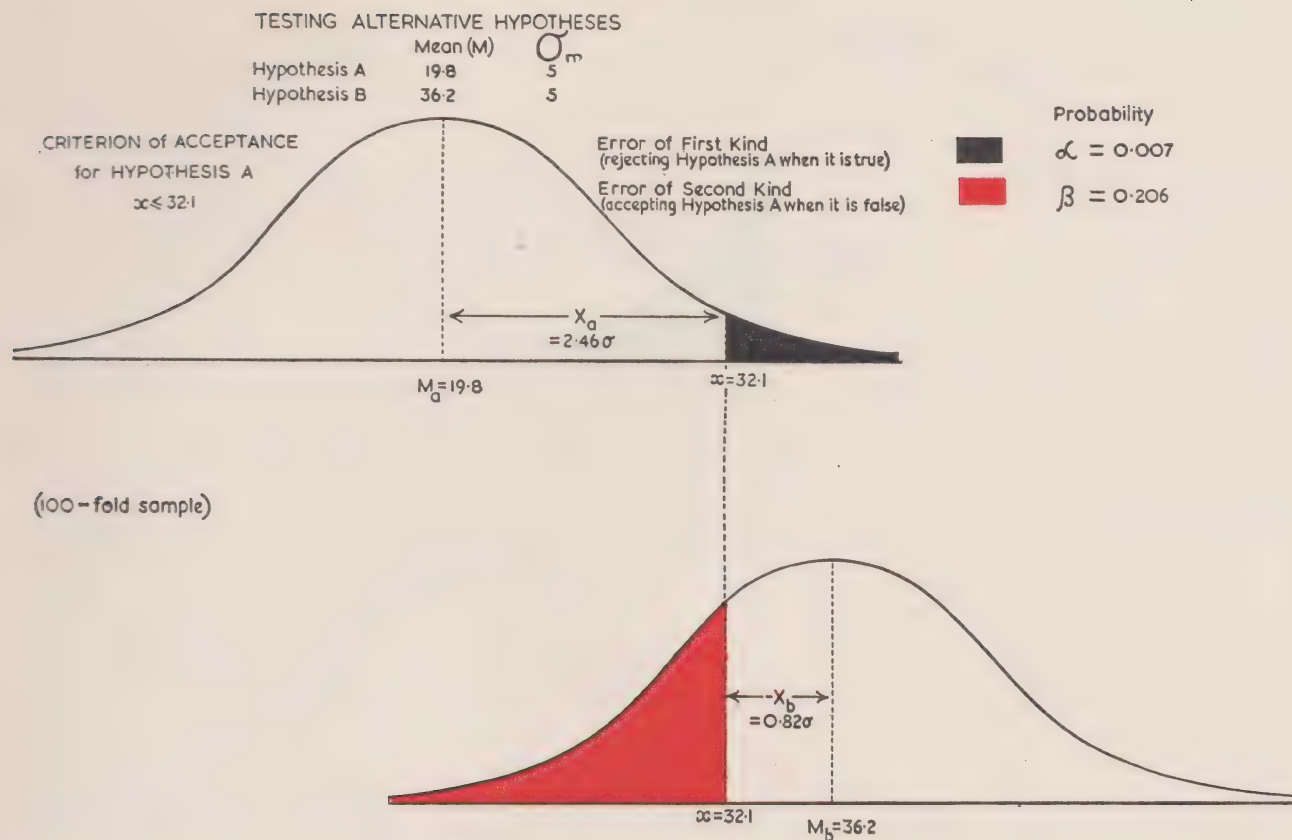


FIG. 133. Testing the same exclusive alternative hypotheses as in Fig. 132 and same sample size, choice of rejection-acceptance criterion which makes the error of the *first* kind smaller, makes the error of the *second* larger.

carriers and 50 normal among 500 female flies in all, i.e.  $P_a = 0.1$  and  $P_b = 0.9$ . In this case  $(\alpha - \beta)P_a = -0.009$ , so that

$$P_t = 0.908 + 0.009 = 0.917 \quad \text{or} \quad 91.7 \text{ per cent.}$$

If the null hypothesis is referable to the *smaller* population at risk (i.e. if it has lower prior probability than the alternative) the effect of making the rejection criterion more exacting is to *lower the probability of arriving at a correct decision*.

Before discussing how far this rule is of general application within the framework of our model assumptions, let us take stock of another highly relevant variable, *viz.* sample size. For a fixed size of sample the foregoing results have sufficiently emphasised what a visual diagram suffices to demonstrate (Figs. 132-134), i.e. we cannot decrease the conditional probability ( $\alpha$ ) of an error of the first kind without increasing the conditional probability ( $\beta$ ) of an error of the second kind and *vice versa*. It is also of importance to appreciate that we can make  $\beta$  for a preassigned value of  $\alpha$  as small as we wish to make it, only if we make the size of the sample large enough. Conversely, we can keep  $\alpha$  at a preassigned level for a smaller test sample only by making  $\beta$  larger.

Consider for example the consequence of applying the foregoing test to fraternities of 100, so that  $M_a = 50$ ;  $\sigma_a = 5$ ;  $M_b = 66.6$  and  $\sigma_b = 4.71$ . If we make the vector criterion of acceptance or rejection conveniently near the  $2\sigma$  level, we shall set it on either side of a score level 60.5 in which event  $(x - M_a)$  is approximately  $2.1\sigma_a$  and  $\alpha \simeq 0.036$ . If so,  $(x - M_b)$  is approximately  $-1.32\sigma_b$  and  $\beta \simeq 0.0934$ . The value of  $\alpha$  here agrees as closely as need be for exemplary purposes with the value chosen ( $\alpha = 0.038$ ) for the 144-fold sample when

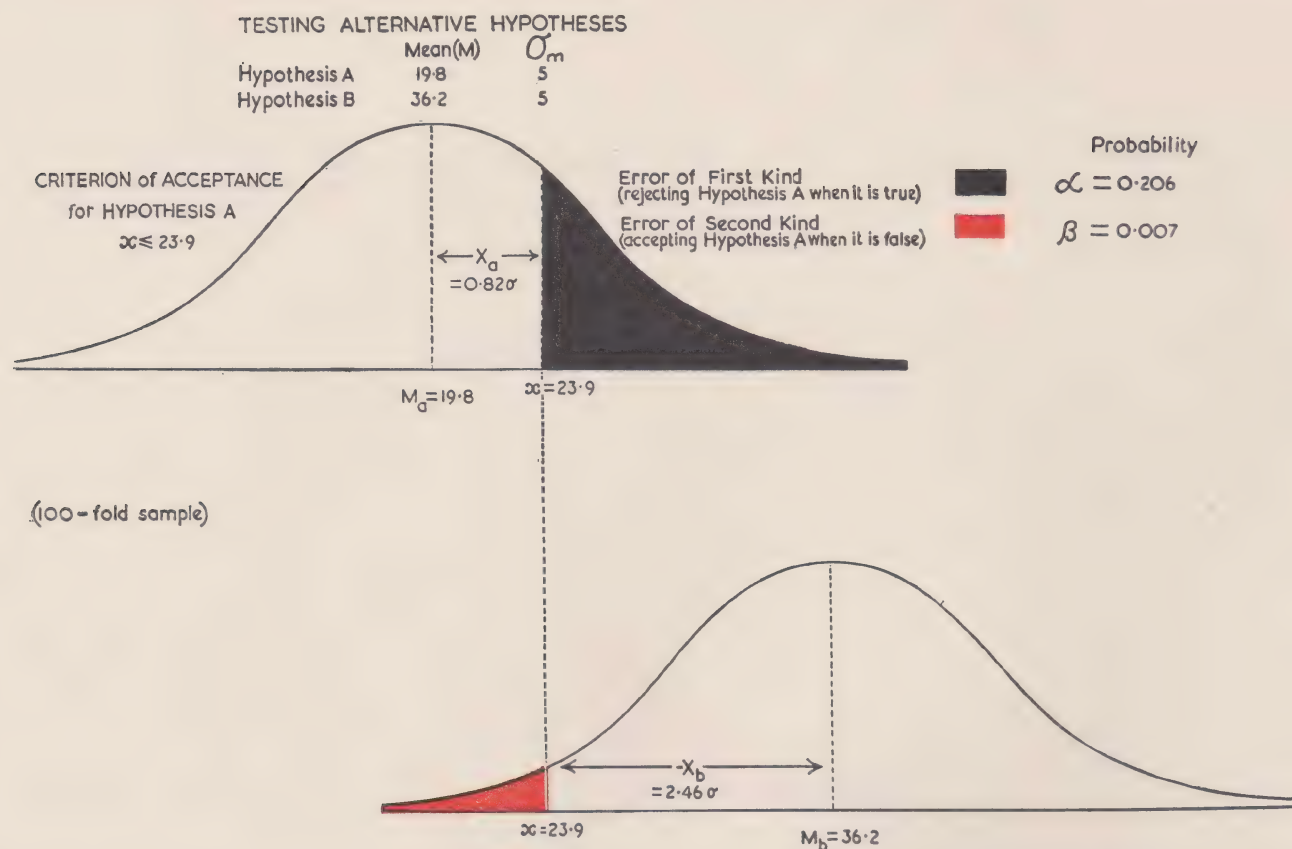


FIG. 134. Testing the same exclusive alternative hypotheses as in Fig. 132 and same sample size, choice of rejection-acceptance criterion which makes the error of the *second* kind smaller, makes the error of the *first* larger.

$\beta \simeq 0.021$ . Thus the effect of reducing the sample size is to increase more than 4-fold the probability of an error of the second kind for a corresponding probability of error of the first kind.

In this case we can make our two error risks nearly equal by setting our limits of rejection and acceptance for the null hypothesis on either side of the score  $x = 58.5$ , in which event the null hypothesis sets the upper limit of acceptance at  $+1.7\sigma_a$  and the alternative sets the lower limit for rejection at  $-1.74\sigma_b$ . Thus  $\alpha \simeq 0.045$  and  $\beta \simeq 0.041$ . If  $P_a = 0.9$  as in our first example  $P_t = 95.9$  per cent. For the 144-fold sample we obtained  $P_t \simeq 98.1$  per cent. when the two conditional risks were nearly equal.

Before we go further, we may well retrace our steps. We made the arbitrary decision to designate as our null hypothesis the assertion that the mother fly is normal. Actually, we have given no reason for doing so; and we may pause at this stage to dispose of a misconception widespread among those who carry out routine tests within the framework of the unique null hypothesis. There is prevalent a somewhat naïve view that we choose our null hypothesis as a safeguard against wishful thinking, and that we make accordingly our criterion of rejection as exacting as need be. On such a view our criterion of rejection is at best a *disciplinary convention*; and as such has nothing to do with *unconditional* statistical inference. Also one can justify it as such only if one chooses the null hypothesis on the understanding that one wishes to fall backwards in preserving one's rectitude, i.e. if the null hypothesis is actually the one the investigator has reasons for believing to be false. Evidently no recipe that the best Mrs. Beeton can prescribe will indeed meet one's requirements in all situations. If experiments on laboratory stocks have convinced the investigator that a new therapy is preferable to a current



procedure, the enthusiastic research worker will not reasonably impose on the null hypothesis a criterion of rejection as exacting as that of the sceptical investigator undertaking experiments to test the credentials of extrasensory perception. In conformity with current procedure, he or she will nevertheless invoke a null hypothesis of the same type in either situation, and with the same convention (e.g. 0.05 significance level) of rejection, if accustomed to rely on current cookery book recipes. The reason is that the cookery book recipe will commonly prescribe as the appropriate null hypothesis the one which commends itself to the mathematician because he can manipulate it algebraically, i.e. for reasons which have nothing to do with the operational intention of the scientific worker.

In the model situation discussed hitherto, we have, in fact, sidestepped the temptation to choose our null hypothesis for this reason, since it would be equally easy to adopt as such the postulate that the fly mother is a lethal carrier. A rejection criterion identical in terms of the conditional risk of error of the first kind, as is indeed the most we can specify within the framework of a unique null hypothesis, would then lead us to results numerically inconsistent with those we have so far explored. The reader may check this assertion by reversing the role of the two hypotheses in the foregoing examples.

Partly because of the size of the samples chosen, previous tests in our model situation have led to a high probability of correct decision arrived at in conformity with traditional procedure, i.e. within the framework of the unique null hypothesis. This may lead us to a totally wrong view of what we can rely on it to accomplish, if we fail to take stock of two background conditions plausibly invoked in the prescribed set-up, but rarely admissible in other situations,

- (a) we concede only one admissible alternative to the null hypothesis;
- (b) we have postulated a complete specification of the sampling distribution in terms of the alternative thereto.

It will be simpler, if we first examine the implications of (b). In all examples hitherto cited we have found that  $P_t \geq 0.5$ , i.e. that more than 50 per cent. of our decisions will be right if we consistently follow the last prescription, in which event we shall be more often right than wrong. Now there is no reason why this should be so, other than the fact that we can here fix in advance the *size of the sample* and the criterion of rejection or acceptance for the null hypothesis with due regard to the value of the relevant parameters of *both* hypotheses. To clarify the relevance of the consideration last stated, let us now replace the female carriers of a sex-linked lethal gene by females with a virus infection to which their male offspring succumb somewhat more readily than their sister flies. We shall postulate a sex ratio of 11 : 9 in favour of females among the progeny of infected mothers. Our alternative hypothesis is now that  $p_b = 0.55$ .

On the new hypothesis which we again provisionally assume to be the only admissible alternative to the null one ( $p_a = \frac{1}{2}$ ), we have  $M_b = 55$  and  $\sigma_b \simeq 4.98$  for fraternities of 100. If we set our rejection criterion on either side of  $x = 60.5$  we have as before  $(x - M_a) = +2.1\sigma$  and  $(x - M_b) = +1.1\sigma$ , whence  $\alpha = 0.018$  and  $\beta = 0.864$ . Thus  $(1 - \beta) = 0.136$  and  $(\alpha - \beta) = -0.846$ , whence from (i)

$$P_t \simeq 89.7 \text{ per cent. when } P_a = 0.9;$$

$$P_t \simeq 22.0 \text{ per cent. when } P_a = 0.1.$$

This example illustrates the important role of the *prior probabilities*  $P_a$  and  $P_b = (1 - P_a)$  which define the *populations at risk under each hypothesis*. If  $P_a = 0.1$  we have  $P_t = 22$  per cent. and  $P_f = 78$  per cent., i.e. consistent application of the rule will lead us to be *wrong more often than right* when the sample is as small as 100, but we can fix the size of the sample to ensure that  $P_t > \frac{1}{2}$  only if we can assign at least a lower limit to  $P_a$ , and then only if we can



assign a value to  $\beta$ . Now we cannot assign a value to  $\beta$  unless we can specify the appropriate parameter (in this case  $p_b$ ) or parameters of the sample distribution of the admissible alternative hypothesis. In any case, it will rarely happen that we as easily conceptualise the meaning we can here confer on  $P_a$ , and still more rarely that we can equip it with a numerical value.

The considerations last stated do not exhaust limitations of current test procedure within the framework of the unique null hypothesis. We have so far assumed a single admissible alternative and we can very rarely, if ever, make such an assumption with propriety. Indeed, much statistical enquiry goes on against a background of an infinitude of admissible alternatives to any null hypothesis we may advance. In our model set-up we have already adumbrated this complication ; and we may now with profit postulate 3 admissible hypotheses. We shall then get into focus a general conclusion concerning the utmost we can legitimately infer in the domain of unconditional assertion from the outcome of a decision test referable to a unique null hypothesis in the absence of background information concerning admissible alternatives thereto. We may anticipate it by reinterpreting (ii) above. When  $\beta > \alpha$  in (ii),  $(\alpha - \beta)$  is negative and  $P_f < \beta$ , i.e. the uncertainty safeguard is *less than the greater conditional risk*. When  $\beta < \alpha$ ,  $(\alpha - \beta)$  is positive and  $P_f > \beta$ , i.e. the uncertainty safeguard is *greater than the smaller conditional risk*.

The conclusions last stated refer to a situation in which only two hypotheses are admissible. An examination of a model situation in which three hypotheses are admissible will exhibit it as a particular case of a general proposition expressible as follows on the assumption that our concern is with the long-run proportion of true assertions we make on the basis of a test procedure of the type under discussion :

- (i) the worst that can happen is that we shall exclusively encounter situations prescribed by the hypothesis associated with the greatest conditional risk ( $\rho_g$ ) of erroneous rejection ;
- (ii) the best that can happen is that we shall exclusively encounter situations prescribed by the hypothesis associated with the smallest conditional risk ( $\rho_s$ ) of erroneous rejection.

Our new model will be that the following three hypotheses are admissible w.r.t. our fruitfly culture ; and we therefore start by assuming that we can define the relevant parameters and sample distributions referable to each of them :

*Hypothesis A:* The female is normal,  $p_a = \frac{1}{2} = \frac{30}{60}$ .

*Hypothesis B:* The female transmits a virus infection,  $p_b = \frac{1}{2} \frac{1}{0} = \frac{3}{6} 0$ ,

*Hypothesis C:* The female carries a sex-linked lethal,  $p_c = \frac{2}{3} = \frac{40}{60}$ .

We shall designate the proportions of the 3 types of female flies as respectively  $P_a$ ,  $P_b$ ,  $P_c$  so that  $(P_a + P_b + P_c) = 1$ . In real life there is no reason to exclude the possibility that a lethal carrier could transmit the virus ; but we assume for argument that such flies are 100 per cent. resistant to it. We are free to select any one of the above as our null hypothesis ; and shall first assume that our null hypothesis is  $p = p_a$ . In that event we may choose a single rejection criterion  $x > x_{bc}$ , thereby fixing the conditional risk of rejecting the null hypothesis when true as  $\alpha$ , that of rejecting hypothesis  $B$  when true as  $\beta$  and that of rejecting hypothesis  $C$  when true as  $\gamma$ . The conditional probability of an error of the first kind is  $\alpha P_a$ . We accept the null hypothesis erroneously, if we either reject  $B$  when true or reject  $C$  when true. Whence the probability of wrong acceptance is  $\beta P_b + \gamma P_c$ , and the probability of making a false decision of one sort or the other is

$$P_f = \alpha P_a + \beta P_b + \gamma P_c \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (iii)$$

This is equivalent to

$$P_t = (1 - P_f) = 1 - \alpha P_a - \beta P_b - \gamma P_c.$$









to make the probability ( $\alpha$ ) of rejecting the null hypothesis when true, so that  $\alpha = \rho_a$ , whence  $P_f < \alpha$ . By appropriate choice of sample size we can then make the probability of a correct verdict *on the null hypothesis* as near to unity as we like without invoking any information w.r.t. its *prior probability*.\* We thus arrive at the following conclusion: *a test procedure may be informative in the domain of unconditional inference if, and only if, we can precisely specify each of an exhaustive and exclusive set of hypotheses.*

The last statement calls for qualification on two counts. A parameter  $p_h$  definitive of an admissible alternative to the null hypothesis ( $p = p_0$ ) may be indefinitely close to  $p_0$  itself. If so (Fig. 135),  $\rho_h \simeq 1 - \rho_0$  for samples of finite size and we can make both  $\rho_h$  and  $\rho_0$  indefinitely

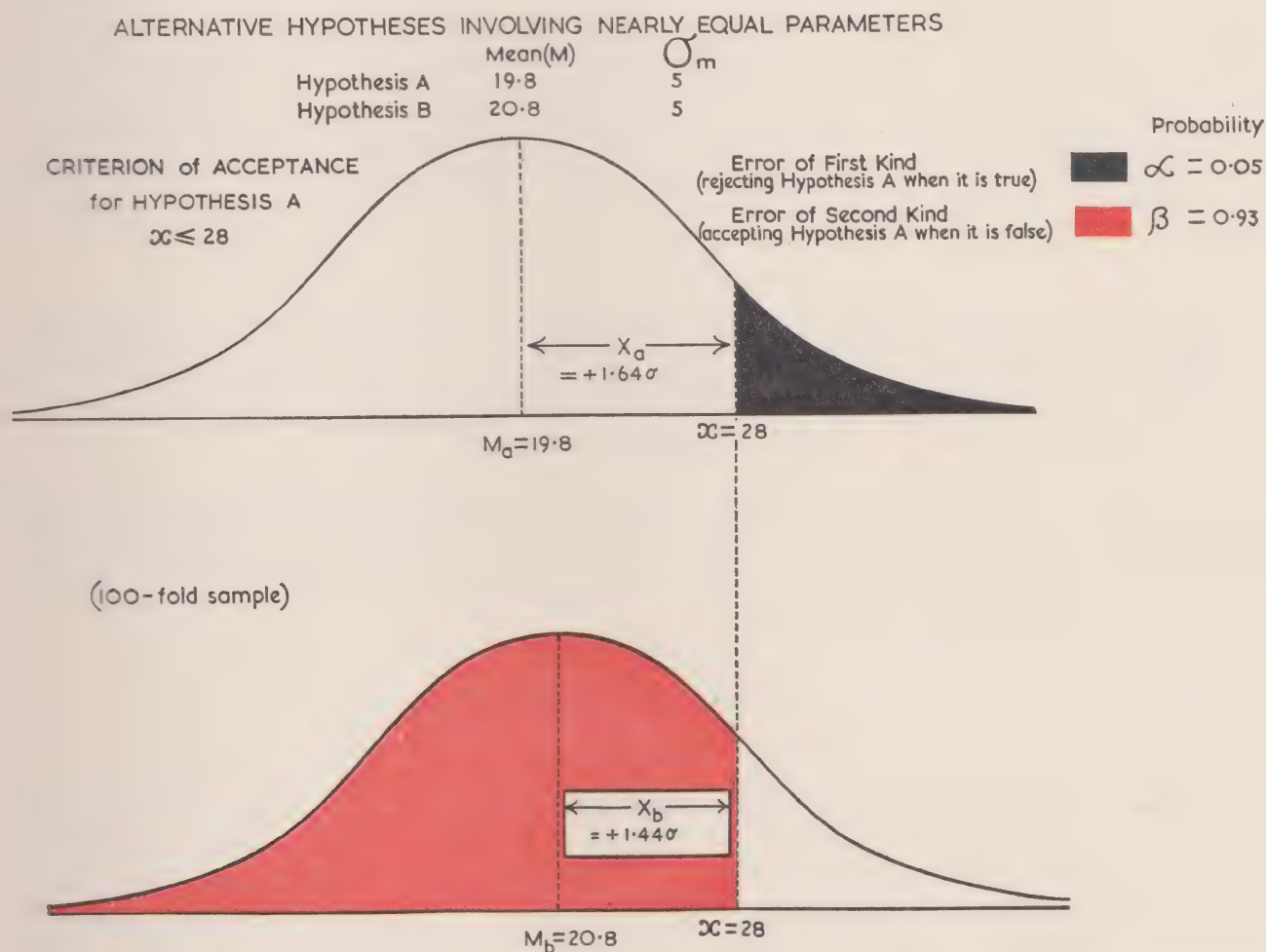


FIG. 135. Testing Exclusive Alternative Hypotheses. The sample size and the variance of the u.s.d. for each hypothesis as in Fig. 132 and the null hypothesis ( $M_a = 19.8$ ) unchanged. By making our alternative hypothesis that  $M_b$  lies very near  $M_a$  without changing the size of the sample, we can make  $\beta \simeq (1 - \alpha)$ .

small only by making our sample indefinitely large. This consideration has an important bearing on the concept of test power touched on below. Here it is relevant because we can rarely be certain that no such hypothesis alternative to the one chosen as null is indeed admissible. This raises a question of pivotal importance in connexion with the foregoing exposition: in what situations can one postulate an exhaustive and exclusive set of admissible hypotheses which fulfil all the relevant conditions now stated?

\* Neyman and Pearson (1933), "Testing Statistical Hypotheses in Relation to Probabilities *a priori*," *Proc. Camb. Phil. Soc.*, Vol. 29.

An important class in which the postulate of an exclusive and exhaustive set of admissible hypotheses is legitimate arises in pathology when we can

- (i) classify test subjects as healthy or sufferers from a particular disease ;
- (ii) assign a probability on the basis of laboratory experience to the assertion that a single test will fail to identify them correctly.

For heuristic purposes a criterion for screening tuberculous patients cited by Neyman will serve as a type specimen.\* On the basis of laboratory experience, we assume that a single X-ray film will: (a) fail to detect the disease in 40 per cent. of sufferers; (b) give a positive result for 1 per cent. of healthy test subjects. Clearly we need to make more than one film, if we aim at a high level of satisfactory diagnosis. We shall assume that we take 5 and adopt as our test criterion the rule that we deem the disease to be present if at least one film is classifiable as positive. Our first which we may call the null hypothesis (*hypothesis A*) will be that the disease is present. Our test criterion leads us to reject the hypothesis if all 5 films are negative. The relevant parameter ( $p_a$ ) is the probability of failure, in this case 0.4. Hence the probability of rejecting the null hypothesis (i.e. wrongly classifying the test subject as healthy) is

$$\alpha = (0.4)^5 = 0.01024.$$

The alternative hypothesis (*hypothesis B*) is that the test subject is healthy. If so, the probability ( $p_b$ ) of getting a negative result from one film is 0.99 and that of getting 5 negative results is  $(0.99)^5$ . We shall reject the alternative hypothesis if at least one film is positive, i.e.

$$\beta = 1 - (0.99)^5 = 0.04901.$$

The reader will find it instructive to explore the outcome of different test criteria based on different sizes of sample (i.e. numbers of films per test subject) within the framework of the foregoing assumptions. In this context  $P_a$  in (i) above is the incidence of tuberculosis in the population. Our truth equation is

$$\begin{aligned} P_t &= 1 - \beta - (\alpha - \beta)P_a \\ &= 1 - 0.04901 - (0.01024 - 0.04901)P_a \\ &= 0.95099 + (0.03877)P_a. \end{aligned}$$

Thus the test criterion ensures a probability of overall correct decision a little more than 95.5 per cent. if the incidence of the disease is 1 per cent., and must inevitably ensure a figure more than 95 per cent. for an overall correct verdict in accordance with (xi) of 20.03. However, we shall now see that unconditional assertions of this sort are of trivial interest in connection with decisions of this kind, though the procedure illustrated may well have applications in the domain of differential diagnosis.†

Throughout our treatment of the model situations explored in this section we have assumed that the fly cultures are composite; and that  $P_a$ ,  $P_b$ , etc., specify existent subpopulations at risk. We have then the model situation to which Bayes' theorem refers. In an actual situation we

\* The medical specialist will recognise some arbitrary assumptions in the argument here advanced for illustrative purposes alone.

† The writer is greatly indebted to Mr. R. Wrighton for suggestions and criticisms incorporated in this and the ensuing section. Since it went to press he has called my attention to two important contributions which lay bare the limitations of the test procedures commonly employed in biological and sociological work. Jackson (*Stat. Res. Mem.* 1, 1936) introduces the concept of *stringency*, a test being most stringent if it assigns a minimal unconditional *uncertainty safeguard* (Wrighton) as here defined. V. Mises (*Ann. Math. Stat.* 14, 238) uses the term *error chance* in the same sense and speaks of the *success rate* of a test as equivalent to  $P_t$  in the foregoing discussion. The term *stochastic credibility* suggests itself as of wider applicability in the common domain of test procedure and estimation.



might not know whether the culture is homogeneous or composite. Alternatively, we might know that the culture contains flies of only one sort without knowing which. From the viewpoint of Bayes' theorem, we might then state that one value of  $P_h$  is unity and every other value of  $P_h$  is zero; but we should know the answer to our problem only if we could identify the hypothesis  $H$  which assigns  $P_h = 1$ . The test procedure for two exclusive alternative hypotheses sidesteps both horns of the Bayes' dilemma. If we know the culture is composite, we need not know the contribution of each population at risk. If we know that it is homogeneous or if we merely know that it may be, the same recipe holds good, since two relations hold good for *all* values of  $P_a$  and  $P_b$  including as a limiting case  $P_a = 0$ ,  $P_b = 1$  or *vice versa*. In situations to which the Bayes' balance sheet is factually irrelevant, each hypothesis being referable to an existent population at risk, and in situations to which it is factually irrelevant in the sense that we cannot realistically conceptualise the sampling process in two stages, it is equally true that our uncertainty safeguard ( $P_f$ ) lies between  $\alpha$  and  $\beta$ , being equal to  $\alpha$  if we design a trial to make  $\alpha = \beta$ .

#### 20.04 THE CONCEPT OF TEST POWER

In the foregoing section we have examined a model situation, *viz.* a fruitfly culture, to throw light on the relevance of test procedure to unconditional inference, i.e. our concern has been to assign a probability to a correct decision for or against a hypothesis. On the assumption that the female deemed to be normal in this context is a new and valuable mutant, we might also formulate our problem in terms of conditional inference. Thus we may wish to curtail both the risk of letting lethal genes accumulate in our stock and the risk of destroying normal stock otherwise available for perpetuating it. Accordingly, we decide to screen our females by setting up a rejection criterion which will set an acceptable limit to the risk incurred in retaining a lethal carrier and an acceptable limit to the risk of losing an otherwise normal female which carries the mutant gene we seek to perpetuate.

We can likewise, and usefully, regard the issue at stake in a diagnostic test such as the one Neyman cites as on all fours with decisions which arise in *quality control*, when the end in view is to ensure against incurring hazards respectively (*vide* 20.02) designated as *producer's risk* and *consumer's risk*. The main preoccupation of the administration in the situation last discussed will in fact have less to do with an overall assessment of correct judgment than with the penalties of making mistakes of two sorts. To classify wrongly a tuberculous person is to deprive him or her of proper treatment. To classify wrongly a healthy person is likely to cause unjustifiable alarm and despondency. A test procedure which prescribes that neither risk exceeds what the authorities regard as admissible therefore satisfies the practical demands of the situation from their viewpoint. We may state these demands explicitly in the form:

- (i) *if* the test subject is tuberculous, the risk of erroneous diagnoses must not exceed  $\alpha_1$ ;
- (ii) *if* the test subject is healthy, the risk of erroneous judgment must not exceed  $\alpha_2$ .

Any unconditional statement we can legitimately make in this context presupposes the possibility of classifying the test subjects exclusively as of one or other type; but the administrative intention does not change, if we postulate that an appreciable number of test subjects are unclassifiable by recourse to any available independent diagnostic criterion. The test need not then lead to consequences embarrassing to authorities content to disclaim responsibility for individuals unless definitely deemed to be healthy or tuberculous. Undoubtedly, there will arise in administration many comparable *costing* situations in which a conscientious claim for limiting the requirements of a test to such conditional assertions is admissible; but the propriety of such a procedure in the domain of scientific research is at least open to debate.



Recent literature on quality control techniques justifies the suspicion that some writers would advance the claim that conditional decision tests of this type are appropriate in the domain of the prophylactic or therapeutic trial. It is therefore pertinent to examine the relevance of the analogy between the end in view of the salesman and that of the research worker approaching a clinical trial against the background of laboratory experiments *in vitro* or on animals. In this situation the investigator will not lightly incur the risk of losing credit for a major discovery nor cheerfully shoulder the risk that subsequent enquiry will discredit his conclusions. If content to follow the practice of the large-scale commercial corporation, he will therefore invoke a test procedure which will set appropriate limits to the risk of wrongly rejecting the alternatives : (i) his own assertion that treatment  $B$  guarantees  $b$  per cent. more cures than treatment  $A$  ; (ii) the assertion of an imaginary critic that treatment  $B$  guarantees only  $a$  per cent. more cures than treatment  $A$ . By all too easy stages, statistical inspection then becomes a recipe for statistical careerism. The investigator and his putative opponent relinquish their proper relation as colleagues in the impersonal pursuit of truth to embrace a convention which safeguards the *amour propre* of each. The decision to make the best of a bad job in this sense involves an ethical issue which is not amenable to arguments likely to win universal assent ; but it carries with it an implication which may well damp the enthusiasm of the convert. This will come into focus, if we here digress to clarify the Neyman-Pearson concept of test *power*.

In the taxonomic domain of the 2-class universe we specify  $\alpha$  and  $\beta$  in the following way for the  $r$ -fold sample when the criterion for rejecting hypothesis  $A$  ( $p = p_a$ ) and hence for accepting hypothesis  $B$  ( $p = p_b$ ) is  $x \geq t$  :

$$\alpha = \sum_{x=t}^{x=r} r_{(x)} p_a^x (1 - p_a)^{r-x} ; \quad 1 - \alpha = \sum_{x=0}^{x=(t-1)} r_{(x)} p_a^x (1 - p_a)^{r-x} \quad . \quad . \quad . \quad (i)$$

$$\beta = \sum_{x=0}^{x=(t-1)} r_{(x)} p_b^x (1 - p_b)^{r-x} ; \quad 1 - \beta = \sum_{x=t}^{x=r} r_{(x)} p_b^x (1 - p_b)^{r-x} \quad . \quad . \quad . \quad (ii)$$

What Neyman calls the power function  $F(p)$  of the test for the same size ( $r$ ) of sample and the same criterion score ( $t$ ) is picturable as the graph of the following function over the range  $p = 0$  to  $p = 1$  :

$$F(p) = \sum_{x=t}^{x=r} r_{(x)} p^x (1 - p)^{r-x} \quad . \quad . \quad . \quad . \quad (iii)$$

It follows that

$$F(p_a) = \alpha \quad \text{and} \quad F(p_b) = 1 - \beta \quad . \quad . \quad . \quad . \quad (iv)$$

For a given value of  $r$  and of  $t$ , the condition that  $\alpha = \beta$  is, of course,

$$F(p_a) = 1 - F(p_b) \quad . \quad . \quad . \quad . \quad (v)$$

Having fixed any criterion for rejection of the null hypothesis ( $A$ ), and having chosen the alternative hypothesis ( $p = p_b$ ), we speak of  $F(p_b)$  as the *power of the test*. This being  $(1 - \beta)$  is the probability of rejecting the null hypothesis when it is false, on the assumption that *the alternative is the only admissible one*. One test prescription is more powerful than another if it has a higher power in this sense for a fixed value of  $\alpha$ , i.e. if it assigns a lower probability to error of the second kind for the same probability of error of the first. If the two test prescriptions both invoke the same distributions, the test which employs a larger sample must be the more powerful one.

The reader will find it instructive to plot  $F(p)$  against  $p$  for the following example of a test procedure. The null hypothesis is that  $p = \frac{1}{2}$  when  $r = 144$ . The rejection criterion is  $x > 82.5$ . For the distribution prescribed by the null hypothesis the mean is 72 with  $\sigma = 6$ . Whence



the criterion score in standard form is  $(82.5 - 72) \div 6 = 1.75$ . This excludes 4 per cent. of the area of the normal fitting curve, i.e.  $\alpha = 0.04$ . For this set-up we may tabulate as below :

TABLE 1

$p$	$M$	$X = (82.5 - M)$	$\sigma$	$c = X \div \sigma$	$\beta$	$F(p) = (1 - \beta)$
$\frac{1.0}{2.4}$	60	22.5	5.9161	3.8032	$> 0.999$	$< 0.001$
$\frac{1.1}{2.4}$	66	16.5	5.9791	2.7596	0.996	0.004
$\frac{1.3}{2.4}$	78	4.5	5.9791	0.7526	0.773	0.227
$\frac{1.4}{2.4}$	84	-1.5	5.9161	-0.2535	0.401	0.599
$\frac{1.5}{2.4}$	90	-7.5	5.8095	-1.2910	0.099	0.901
$\frac{1.6}{2.4}$	96	-13.5	5.6568	-2.3865	0.013	0.987
$\frac{1.7}{2.4}$	102	-19.5	5.4544	-3.5750	$< 0.001$	$> 0.999$

The concept of test power is easily interpretable in the alternative domain of representative scoring. The simplest type of test is then one which invokes alternative hypotheses specifying the mean score of the u.s.d. for each of two normal universes. Consider now the following model. We do not know the mean value ( $M$ ) of the normal variate ; but we do know that the variance of the u.s.d. is 2500. Whence that  $(\sigma_m^2)$  of the mean of the  $r$ -fold sample is  $2500 \div r$ . For the 100-fold sample  $\sigma_m = 5$ . Our null hypothesis is that  $M = 18.2$ . The standard score corresponding to a sample value ( $M_x$ ) of the mean is therefore  $(M_x - 18.2) \div 5$ . To make  $\alpha = 0.05$  we must make the deviation equal to  $1.64\sigma_m$ , i.e.  $(M_x - 18.2) = 8.2$ , whence  $M_x = 26.4$ .

The alternative hypothesis which makes  $\beta = 0.05$  is that  $(M_x - M) \div \sigma_m = -1.64$ , so that  $(26.4 - M) = -5(1.64)$ . Whence the hypothesis is that  $M = 26.4 + 8.2 = 34.6$ . If our alternative hypothesis were that  $M = 28.2$ , the score deviation would be  $(26.4 - 28.2) = -1.8$  or  $-0.36\sigma_m$ . At this level  $\beta = 0.359$ . To make the two risks equal when the sample size is 100 and the alternative hypothesis is  $M = 28.2$  we must choose the sample value ( $M_x$ ) definitive of our rejection criterion, so that

$$\frac{M_x - 18.2}{5} = - \frac{(M_x - 28.2)}{5}.$$

In this case  $M_x = 23.2$ , i.e.  $(M_x - 18.2) = 5 = \sigma_m$ , so that  $\alpha = 0.159 = \beta$ . To equalise both risks as nearly as possible at the level  $\alpha = 0.05 = \beta$ , when the alternative hypothesis is that  $M = 28.2$ , we must enlarge our sample size ( $r$ ) so that  $\sigma_m = 50 \div \sqrt{r}$  in the identity

$$\frac{23.2 - 18.2}{\sigma_m} = 1.64 = \frac{-(23.2 - 28.2)}{\sigma_m}.$$

Whence we get

$$\sqrt{r} = 16.4.$$

Whence  $r = 269$  to the nearest integer.

We may generalise the rules of test prescription thus :

(i) to fix  $\alpha$  at  $h_a\sigma_m$  level we make our test criterion

$$\frac{(t - M_a)}{\sigma_m} = h_a \quad \text{so that} \quad t = M_a + h_a\sigma_m \quad . \quad . \quad . \quad (vi)$$

(i) Value of $M$ definitive of alternative hypothesis	(ii) Level of rejection ( $h$ ) expressed as $h\sigma_m$	(iii) Corresponding value of $\beta$	(iv) Power of test criterion ( $1 - \beta$ )	(v) Value of $\beta$ when $\alpha = \beta$	(vi) Value of $r$ when $\alpha = 0.05 = \beta$
18.4	—1.6	0.945	0.055	0.492	672,400
19.4	—1.4	0.919	0.081	0.452	18,678
22.4	—0.8	0.788	0.212	0.337	1,525
24.4	—0.4	0.655	0.345	0.268	700
26.4	0.0	0.500	0.500	0.206	400
28.4	0.4	0.345	0.655	0.154	259
30.4	0.8	0.212	0.788	0.111	181
32.4	1.2	0.115	0.855	0.078	133
34.4	1.6	0.055	0.945	0.053	102
36.4	2.0	0.023	0.977	0.034	81
38.4	2.4	0.008	0.992	0.022	66
40.4	2.8	0.005	0.995	0.013	55



Since the size of sample fixes the power of the test for a fixed value of  $\alpha$ , we can set  $\beta$  at a predestined level appropriate to any single chosen alternative of the null hypothesis only if we plot  $P_f$  for different values of  $r$ . Table 3 sufficiently illustrates the procedure; and the reader may find it instructive to check the arithmetic by recourse to the foregoing equations.

TABLE 3

*Power function ( $P_f = 1 - \beta$ ) for the same model as in Table 1, tabulated separately for different values of sample size  $r$ , with the same rejection criterion ( $\alpha = 0.05$ ) for the null hypothesis ( $M = 18.2$ ), when  $\sigma_m = 5$  for the 100-fold sample. As the head of the columns are score values ( $M_x$ ) corresponding to the condition  $\alpha = 0.05$ , and values of  $\sigma_m$  for the appropriate value of  $r$ .*

Hypothesis $M =$	Size of Sample			
	81	144	256	324
	$\sigma_m = 5.5$ $M_x = 27.3$	$\sigma_m = 4.16$ $M_x = 25.03$	$\sigma_m = 3.125$ $M_x = 23.33$	$\sigma_m = 2.6316$ $M_x = 22.52$
19	0.056	0.073	0.082	0.090
21	0.129	0.166	0.227	0.281
23	0.221	0.312	0.456	0.571
25	0.341	0.496	0.702	0.826
27	0.480	0.681	0.879	0.955
29	0.622	0.829	0.965	0.993
31	0.749	0.924	0.993	0.999
33	0.849	0.972	0.999	0.999
36	0.942	0.996	0.999	0.999
39	0.983	0.999	0.999	0.999
41	0.993	0.999	0.999	0.999

We are now in a position to see more clearly the implications of approaching the interpretation of the outcome of a prophylactic or therapeutic trial as one of accommodating the producer's risk and the consumer's risk in the theory of quality control. If we do so we conceive the test procedure as a game of chance in which the investigator arranges the stakes to accommodate the inclinations of a wholly imaginary contestant. His assertion is that treatment  $B$  guarantees  $b$  per cent. more cures than treatment  $A$ , and this fixes the value of  $b$ . His fictitious opponent asserts that treatment  $B$  guarantees only  $a$  per cent. more cures; but because his opponent is merely a figment of his own fears, all that he can say about  $a$  is that:  $a < b$ . If he conceives that his opponent is ready to deny any operational advantage ( $a = 0$ ) for treatment  $B$ , he may set his own risk as equal to that of his opponent at a much lower level for a fixed size of sample than will be possible if his opponent makes a far more moderate claim (e.g.  $a = \frac{1}{4}b$ ). Alternatively, the design of a test to equalise risks at one and the same level will prescribe the availability of larger samples, if he conceives that his opponent, having first denied any advantage, will subsequently concede that there is some.

Having chosen the form of his own assertion (here the numerical value of  $b$ ), the only exclusive admissible alternative with which he can equip the imaginary disputant of his claim is  $a < b$ ; but the alternative test procedure then prescribes recourse to an *infinite sample* as a prerequisite to a firm decision. Within the restricted framework of conditional inference, the alternative test procedure can thus offer no simple nor unique recipe for validifying the operational advantage claimed for a new procedure. Should the reader find the foregoing argument obscure, it may help to clarify the issue, if we go back to the data of Tables 2 and 3 (in which  $\sigma_m = 5$  when  $r = 100$ ). We shall suppose that: (i) disputant  $A$  initially asserts that  $M = 18.2$  and disputant  $B$  initially asserts that  $M = 26.4$ ; (ii) both disputants initially



agree to accept the outcome of a test which vindicates the claim of  $B$  if the 400-fold sample value of  $M_x$  exceeds 22.3. In that event each takes a 5 per cent. risk of being discredited. We now suppose that an arbitrator persuades disputant  $A$  to concede that  $M = 19$  and disputant  $B$  to concede that  $M = 25.6$  still taking equal risk on the outcome of a 400-fold trial. Disputant  $A$  will still lose his case if  $M_x > 22.3$ , but each disputant now incurs a 26.8 per cent. risk ( $\alpha = 0.2676 = \beta$ ) of getting an adverse verdict.

From the foregoing heuristic discussion the reader must not infer that the quality control procedures permit us to assign uncertainty safeguards only to conditional assertions. This is not so, as we shall now see. Throughout this section we have provisionally presumed a distinction between conditional and unconditional assertions in terms of the uses to which we put them. This is clear-cut in the sense that : (a) any statement worthy to rank in the corpus of scientific knowledge is one which we can rightly describe as unconditional in the sense elsewhere defined ; (b) statements of the conditional sort suffice as a basis for administrative decision. It is none the less possible to formulate rules of decision leading to unconditional statements of a sort rarely, if ever, relevant to the domain of research in pure science and no more useful to the administrator because more comprehensive in scope than a corresponding statement expressed in the more restricted form. Further consideration of the *Drosophila* model of 20.03 will make this clear.

In 20.03 we set up two hypotheses  $H_a$  that  $p = \frac{1}{2} = p_a$  and  $H_b$  that  $p = \frac{2}{3} = p_b$ ,  $p$  being the probability that any offspring of a particular mother will be female. If we make the rule to reject  $H_a$  if  $x > a + \frac{1}{2}$  for the  $r$ -fold sample and denote by  $L_{x \cdot a}$  the probability that it will contain  $x$  females if  $p = p_a$ , we may assign as the risk ( $\alpha$ ) of rejection when  $H_a$  is true :

$$\alpha = \sum_{x=(a+1)}^{x=r} L_{x \cdot a}.$$

Similarly we may adopt  $H_b$  as our null hypothesis and make the rule to reject it if  $x < b + \frac{1}{2}$ . The corresponding conditional risk ( $\beta$ ) of rejection is then

$$\beta = \sum_{x=0}^{x=b} L_{x \cdot b}.$$

In either case, we attach an uncertainty safeguard ( $\alpha$  or  $\beta$ ) to a statement which is conditional in the sense that it refers to a risk we take of being wrong if a particular hypothesis is correct. Unless  $a = b$  the simultaneous application of the two tests will not necessarily lead to a decision in favour of either hypothesis ; but we can formulate a composite rule which must do so in the form : reject  $H_a$  if  $x > k + \frac{1}{2}$  and reject  $H_b$  if  $x < (k + \frac{1}{2})$  ; and we may be able to choose  $k$  so that  $\alpha \simeq \gamma \simeq \beta$ , if  $r$  is fairly large. This leads to a conditional assertion which assigns  $\gamma$  as the risk that we shall reject either hypothesis if true ; but we cannot assign an acceptable safeguard to any unconditional assertion about the outcome unless the hypotheses so stated constitute an exhaustive and exclusive set. We can make a more comprehensive type of statement if we restate our hypotheses in the form  $H_a$  that  $p \leq p_a$  and  $H_b$  that  $p \geq p_b$  ; and may still guarantee the termination of the test in a firm decision, if we follow the same composite rule of rejecting  $H_a$  when  $x > (k + \frac{1}{2})$  and rejecting  $H_b$  when  $x < (k + \frac{1}{2})$ . We then define  $L_{x \cdot a}$  and  $L_{x \cdot b}$  in terms of  $p_a$  and  $p_b$  as before ; and fix  $k$  so that

$$\sum_{x=(k+1)}^{x=r} L_{x \cdot a} = \alpha = \sum_{x=0}^{x=k} L_{x \cdot b}.$$

Any value  $p < p_a$  then makes the conditional risk of rejecting  $H_a$  in its new form less than  $\alpha$  ; and any value of  $p > p_b$  makes the conditional risk of rejecting  $H_b$  in its new form less than  $\alpha$ , which we may assign at any acceptable level if free to prescribe the sample size ( $r$ ) in advance,



as in 20.01. The rule itself limits our positive statements to the alternatives  $p > p_a$  and  $p < p_b$ . It prohibits any statement about *an interval in which  $p$  lies*, e.g. any statement of the form  $p_a \leq p \leq p_b$ . To any assertion it does entitle us to make we may attach an uncertainty safeguard  $P_{f \cdot n} \leq \alpha$ . Since this sets a limit to the probability assignable to any false statement we may make, we are entitled to say that  $P_f \leq \alpha$  unconditionally defines the uncertainty safeguard of the entire class of statements which the rule subsumes; but we can state this only because the rule subsumes no *simultaneous* statement concerning the relation of  $p$  to both  $p_a$  and  $p_b$ .

If we know that the *Drosophila* culture contains several different genotypes to which we can assign values of  $p$ , we can meaningfully postulate prior probabilities referable to existent populations at risk to formalise the unconditional character of the final statement which the rule endorses. We must do so with due regard to its content, *viz.* the probability of wrongly rejecting the hypothesis  $p_a < p < p_b$  is zero, since the rule does not allow us to reject it. We may then set out the argument in terms of the following symbols,  $\epsilon$  being positive:

Hypothesis	Prior Probability	Conditional Uncertainty Safeguard
1. $p < p_a$	$P_1$	$P_{f \cdot 1} = \alpha - \epsilon_1$
2. $p = p_a$	$P_2$	$P_{f \cdot 2} = \alpha$
3. $p_a < p < p_b$	$P_3$	$P_{f \cdot 3} = 0$
4. $p = p_b$	$P_4$	$P_{f \cdot 4} = \alpha$
5. $p > p_b$	$P_5$	$P_{f \cdot 5} = \alpha - \epsilon_5$

These hypotheses constitute an exclusive set among which we can accept only one. Hence the addition rule applies, and our unconditional uncertainty safeguard is

$$\begin{aligned}
 P_f &= P_1 \cdot P_{f \cdot 1} + P_2 \cdot P_{f \cdot 2} + P_3 \cdot P_{f \cdot 3} + P_4 \cdot P_{f \cdot 4} + P_5 \cdot P_{f \cdot 5} \\
 &= P_1(\alpha - \epsilon_1) + P_2 \cdot \alpha + P_4 \cdot \alpha + P_5(\alpha - \epsilon_5) \\
 &= (1 - P_3)\alpha - P_1 \cdot \epsilon_1 - P_5 \cdot \epsilon_5 \\
 \therefore P_f &\leq \alpha.
 \end{aligned}$$

The prescription of such a rule presupposes two target values of  $p$  which we can readily conceive in relation to standards of quality and to costing limits in an executive set-up, but the unconditional form the terminal statement assumes when we formulate a rule in this way embodies no relevant information other than the content of two types of conditional assertion. What the choice of the single rejection-acceptance criterion  $k$  accomplishes is that the inspection plan itself achieves its task, i.e. the test must lead to the decision to reject either  $H_a$  or  $H_b$ . In fact, both hypotheses may be wrong; and the unconditional form of the assertion is realisable only because the test can never lead to a corresponding assertion, i.e. a statement to the effect  $p_a < p < p_b$ .

If we operate within the framework of a single hypothesis stated in the form  $p \leq p_a$  or  $p \geq p_b$ , and have defined our rejection criterion so that  $P_f \leq \alpha$  is the probability of rejecting it when true, we are free to limit our verdicts, as R. A. Fisher does indeed prescribe, to the alternatives: hypothesis *false* and hypothesis *unproven*. In the sense that  $P_f \leq \alpha$  is then the probability of erroneously making an allowably *decisive* assertion, we might admittedly say that  $P_f \leq \alpha$  is the unconditional safeguard of our test procedure. We then evade the Neyman-Pearson error of the second kind by exposing ourselves to situations in which the overwhelming majority of our decisions will assign the verdict unproven to a false *null* hypothesis. We can

indeed avoid doing so only by prescribing sample size with due regard to the Neyman-Pearson concept of test power ; but any attempt to rehabilitate the Yule-Fisher significance test on such terms undermines previous claims concerning reliability of inference referable to small samples. We shall examine the implications of the last statement more fully in 20.08.

### 20.05 A SEQUENTIAL TEST PROCEDURE

We are now in a position to reinterpret the issue raised at the end of 20.02, i.e. the interpretation of Sequential Ratio limits ; and shall do so with special reference to the type of alternative test procedure called the *double dichotomy* due to Wald. It will help us to understand its rationale if we first briefly examine a unique null hypothesis test which employs the same principle. Wald's non-sequential so-called *exact test* presupposes that we can pair off unit samples from each of two universes which may or may not have the same composition w.r.t. a particular criterion of classification. For instance, we may take cards with replacement one at a time from each of two packs I and II, not necessarily complete, setting out the results as below :

	From I	From II	No. of pairs	Total
Concordances	Red	Red	30	46
	Black	Black	16	
Discordances	Red	Black	12	36
	Black	Red	24	

The null hypothesis is that the two packs I and II are identical w.r.t. colour composition. If so, the two possible types of discordant pairs must occur with equal frequency in the long run. If pack II contains a higher proportion of red cards, pairs of types  $B - R$  should occur more often than pairs of type  $R - B$ . Accordingly, we disregard the concordant pairs, adopting as our null hypothesis that the unit trial probabilities  $p$  of getting  $B - R$  and  $q$  of getting  $R - B$  are equal, i.e.  $p = \frac{1}{2} = q$ . The expected value for a sample of 36 is therefore 18 for each type. If  $p = \frac{1}{2} = q$  the variance of the raw-score distribution of the 36-fold sample of discordant pairs is  $36 \times \frac{1}{2} \times \frac{1}{2} = 9$  so that  $\sigma = 3$ . Now the deviation of the observed number of  $B - R$  pairs from its expected value is  $24 - 18 = 6$ . To evaluate the probability that the deviation will be as great as  $+6$  we require the area of the histogram in the range from 0 to 23. For samples of 16 or over the normal curve of unit variance gives a good quadrature approximation (a real fit) for the terms of the binomial  $(\frac{1}{2} + \frac{1}{2})^r$  if we make the half interval correction. Thus our concern is with the area of the normal curve of unit variance from  $-\infty$  to

$$\frac{(23\frac{1}{2} - 18)}{3} = 1.83.$$

The table of the normal integral shows that the area so defined is 0.966. *If the null hypothesis is true* the odds are therefore about 30 : 1 against getting a score of 24 or more  $B - R$  pairs in a 36-fold sample of discordant pairs. If we wish to test whether  $m$  discordant pairs of a given type exceed expectation  $(\frac{1}{2}r)$  significantly, we may cite the exact probability of getting a score of  $m$  or more as

$$\sum_{x=m}^{x=r} r_{(x)} \left(\frac{1}{2}\right)^x \left(\frac{1}{2}\right)^{r-x} = \frac{r!}{2^r} \sum_{x=m}^{x=r} \frac{1}{x! (r-x)!}.$$



$$c_m = \frac{m - \frac{1}{2} - \frac{1}{2}r}{\frac{1}{2}\sqrt{r}} = \frac{2m - 1 - r}{\sqrt{r}}.$$

*Double Dichotomy Sequential Test.* Wald's non-sequential test last described is a test which involves a unique null hypothesis. A sequential test based on the same method of classifying the data, i.e. rejection of all concordant pairs, presupposes a hypothesis ( $p_b = mp_a = \frac{1}{2}m$ ) alternative to the null hypothesis ( $p_a = \frac{1}{2}$ ). If we designate discordant pairs of two series  $A$  and  $B$  respectively as successes ( $- +$ ) and failures ( $+ -$ ) our alternative hypothesis is then that the probability of success so defined is  $\frac{1}{2}m$ . Since we can label our pairs at choice we may formulate the alternative on the assumption that successes are more common than failures, i.e.  $1 < m \leq 2$ .

$$\frac{p_a^x \cdot q_a^{r-x}}{p_b^x \cdot q_b^{r-x}} = S_r = \frac{1}{m^x(2-m)^{r-x}}.$$

either  $S_r \geq A$  in which event we accept the null hypothesis ;  
or  $S_r \leq B$  in which event we reject the null hypothesis.

accepting the null hypothesis if  $x \leq a$   
 rejecting the null hypothesis if  $x \geq b$ .

$$S_r = \frac{p_a^a q_a^{r-a}}{p_b^a q_b^{r-a}} = \frac{1}{m^a(2-m)^{r-a}} = A \quad . \quad . \quad . \quad . \quad . \quad (i)$$
[illegible]

We may make the alternative hypotheses more comprehensive with a view to an unconditional form of statement *en rapport* with the argument of pp. 878-9 above, if we put  $p_a \leq \frac{1}{2}$  and  $p_b \geq \frac{1}{2}m$  without affecting the choice of  $a$  and  $b$  implicit in the rejection criteria  $S_r = A$  and  $S_r = B$ . For a given sample size ( $r$ ) we may likewise define  $a$  and  $b$  in terms of the conditional risk

$(P_{f.a} \leq \alpha)$  of erroneously rejecting and the conditional risk  $(P_{f.b} \leq \beta)$  of erroneously accepting the null hypothesis, i.e. rejecting the alternative when true. Thus we may write

$$\alpha = \sum_{x=(a+\frac{1}{2})}^{x=r} r_{(x)} p_a^x q_a^{r-x} \quad \text{and} \quad \beta = \sum_{x=0}^{x=(b-\frac{1}{2})} r_{(x)} p_b^x q_b^{r-x}. \quad . \quad . \quad . \quad (\text{iii})$$

When  $p_a = \frac{1}{2}$  and  $p_b = \frac{1}{2}m$

$$\alpha = 2^{-r} \sum_{x=(a+\frac{1}{2})}^{x=r} r_{(x)} \quad \text{and} \quad \beta = 2^{-r} \sum_{x=0}^{x=(b-\frac{1}{2})} r_{(x)} m^x (2-m)^{r-x};$$

$$(1-\alpha) = 2^{-r} \sum_{x=0}^{x=(a+\frac{1}{2})} r_{(x)} \quad \text{and} \quad (1-\beta) = 2^{-r} \sum_{x=(b-\frac{1}{2})}^{x=r} r_{(x)} m^x (2-m)^{r-x}.$$

Since  $\alpha$  and  $\beta$  suffice to define  $x = a$  and  $x = b$  for the value of  $r$  at which the test terminates, we may hope to fix  $S_r = A$  and  $S_r = B$  to ensure that the unconditional uncertainty safeguard ( $P_f$ ) for a decision one way or the other exceeds preassigned values of neither  $\alpha$  nor  $\beta$ . The fact that the test may terminate when  $p_a < p < p_b$  is then irrelevant, since our verdicts are of only two sorts,  $p \leq p_a$  or  $p \geq p_b$ . The clue to the relation we seek is that: (i)  $\beta$  and  $(1 - \alpha)$  are each referable to advancing totals of a binomial summation, i.e. from 0 up to  $b$  or from 0 up to  $a$ ; (ii)  $\alpha$  and  $(1 - \beta)$  are each referable to receding totals, i.e. from  $r$  down to  $(a + 1)$  or from  $r$  down to  $(b + 1)$ . We shall now state, but without proof at this stage, a rule connecting the sequential ratio with the two classes of totals: (a) the ratio of corresponding *advancing* total frequencies is greater than or equal to the *terminal* sequential ratio ( $p_a^x q_a^{r-x} \div p_b^x q_b^{r-x}$ ) of the sequence; (b) the ratio of corresponding receding total frequencies is always less than or equal to the *initial* sequential ratio of the sequence. The following numerical example for the 4-fold sample distribution when  $p_a = \frac{1}{2}$  and  $p_b = \frac{3}{4}$  illustrates the meaning of this assertion,  $y_a$  and  $y_b$  being relative frequencies on the same scale ( $4^4$ ):

	$x =$	0	1	2	3	4	
$y_a(p_a = \frac{1}{2})$	.	.	16	64	96	64	16
$y_b(p_b = \frac{3}{4})$	.	.	1	12	54	108	81
<i>Advancing Sum</i>		16	80	176	240	256	
<i>Totals</i>		1	13	67	175	256	
<i>Receding Sum</i>		256	240	176	80	16	
<i>Totals</i>		256	255	243	189	81	
<i>Ratio of Advancing Totals</i>		16	6.2	2.6	1.4	1	
<i>Ratio of Receding Totals</i>		1	0.94	0.72	0.42	0.20	
<i>Sequential Ratio</i>	.	.	16	5.3	1.8	0.59	0.20

From (i) and (ii) above we therefore see that

[illegible]

To make the meaning of these inequalities clear, let us first notice that diminishing either  $\alpha$  or  $\beta$  makes the left hand ratio larger and the right hand ratio smaller, e.g.



	$\alpha = 0.10$ $\beta = 0.05$	$\alpha = 0.05$ $= \beta$	$\alpha = 0.05$ $\beta = 0.025$	$\alpha = 0.025$ $= \beta$
$\frac{1-\alpha}{\beta} =$	18	19	38	39
$\frac{\alpha}{1-\beta} =$	$\frac{2}{19}$	$\frac{1}{19}$	$\frac{2}{39}$	$\frac{1}{39}$

Let us now fix  $A = 20$  as our acceptance criterion for the hypothesis  $p \leq p_a$  and  $B = \frac{1}{20}$  as our alternative criterion for acceptance of the hypothesis  $p \geq p_b$ . In accordance with (iv), we have then said that

$$\frac{1-\alpha}{\beta} \geq 20 \quad \text{and} \quad \frac{\alpha}{1-\beta} \leq \frac{1}{20}.$$

We see from the above that this implies that  $\alpha$  and  $\beta$  must then both be less than 0.05. By tabulating  $(1-\alpha)\beta^{-1}$  and  $\alpha(1-\beta)^{-1}$  for descending values of  $\alpha$  and descending values of  $\beta$  we can thus see what sequential ratio limits are consistent with assigning our conditional risks at prescribed levels  $P_{f,a} \leq \alpha$  and  $P_{f,b} \leq \beta$ . Since the test permits us to make only two sorts of statement the overall risk of erroneous decision is  $P_f < \alpha$  if  $\alpha \geq \beta$  and  $P_f < \beta$  if  $\beta \geq \alpha$ .

A proof of the rule stated w.r.t. ratio of advancing and receding totals for 2 binomial series is elementary; and it will suffice to consider the case which arises when  $p_a = \frac{1}{2}$  and  $p_b = \frac{1}{2}m$  in which  $1 < m < 2$ . In this case, we may put  $m = (1+e)$  in which  $e$  is positive, so that

$$k = \frac{m}{2-m} = \frac{1+e}{1-e} > 1.$$

Our acceptance criterion is

$$S_a = \frac{1}{m^a(2-m)^{r-a}}.$$

The ratio of the frequencies of the sum of scores less than or equal to  $a$  is

$$R_a = \frac{\sum_{x=0}^{x=a} r_{(x)}}{\sum_{x=0}^{x=a} r_{(x)} \cdot m^x(2-m)^{r-x}}.$$

We require to prove that  $R_a \geq S_a$  and if  $R_a = K \cdot S_a$  that  $K \geq 1$ . This implies that

$$\begin{aligned} \frac{\sum_{x=0}^{x=a} r_{(x)} \cdot m^a(2-m)^{r-a}}{\sum_{x=0}^{x=a} r_{(x)} \cdot m^x(2-m)^{r-x}} &= K, \\ \therefore K &= \frac{\sum_{x=0}^{x=a} r_{(x)} \cdot m^a(2-m)^{-a}}{\sum_{x=0}^{x=a} r_{(x)} \cdot m^x(2-m)^{-x}} = \frac{\sum_{x=0}^{x=a} r_{(x)} \cdot k^a}{\sum_{x=0}^{x=a} r_{(x)} \cdot k^x}. \end{aligned}$$

Now every value of  $x$  in the denominator of the expression on the right is less than or equal to  $a$ . Since then  $k > 1$ , corresponding terms in the numerator are greater than those of the

denominator, whence  $K > 1$ . The reader should be able to complete the proof for receding totals in the same way. The generalisation of the rule for  $p_b > p_a$  when  $p_a$  is not equal to 0.5 is trivial, since the relevant ratios are then expressible in terms of  $p_b = \frac{1}{2}m_b$  and  $p_a = \frac{1}{2}m_a$ , in which  $m_b > m_a$ .

Let us now return to our test prescription, *viz.* :

- (i) continue to enlarge the sample if  $A > S_r > B$ ;
- (ii) when  $S_r \geq A$  terminate the test by accepting the null hypothesis;
- (iii) when  $S_r \leq B$  terminate the test by rejecting it.

We have elsewhere (20.01 and 20.04) offered alternative interpretations of  $A$  and  $B$ . If we set  $\alpha = 0.05 = \beta$  our choice will be  $A = 20$  and  $B = 0.05$ . The procedure is then reducible to rule of thumb. If  $S_r \geq A$

$$m^x(2-m)^{r-x} \leq \frac{1}{A},$$

$$\therefore x \log m + (r-x) \log (2-m) \leq -\log A,$$

$$\therefore x \leq \frac{\log A + r \log (2-m)}{\log (2-m) - \log m} \quad \dots \quad (v)$$

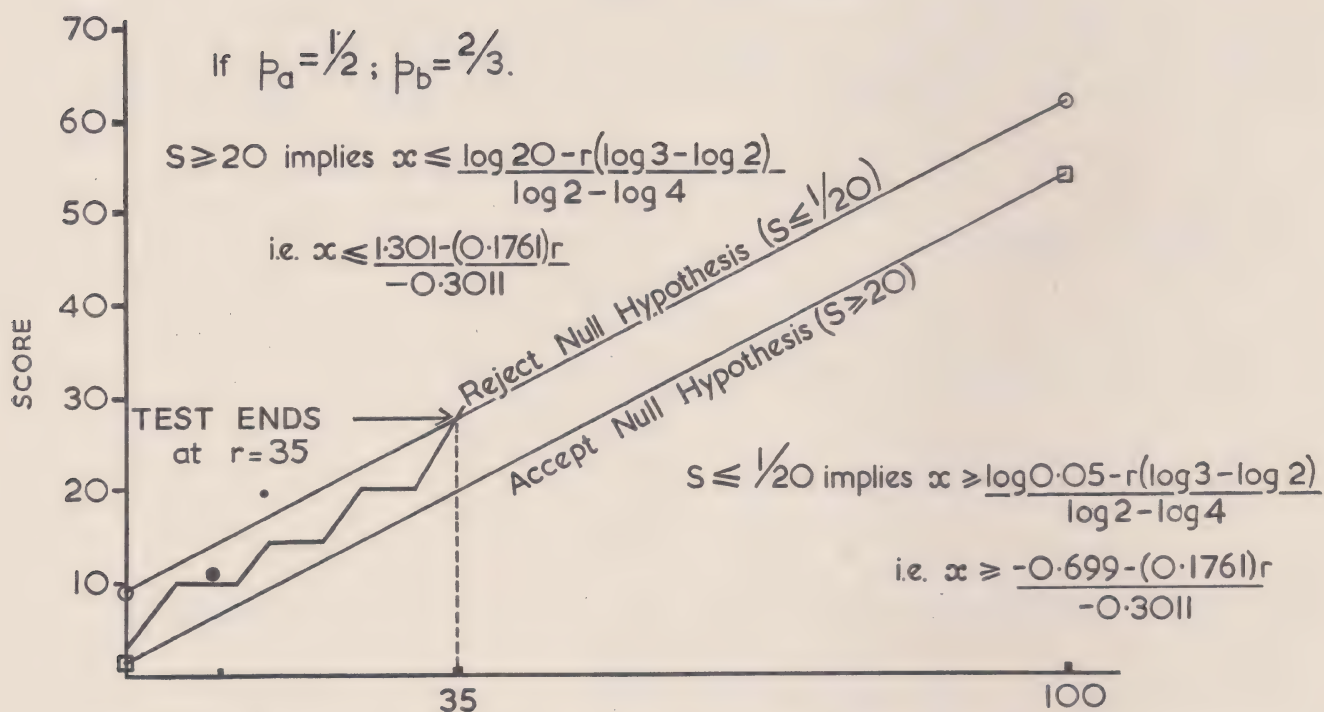


FIG. 136. The Sequential Ratio Test Prescription for Double Dichotomies. The limiting ratios are 9 (9 : 1) and 0.1 (1 : 9) as in the example.

We can make a graph (Fig. 136) of values the *linear* expression on the right of (v) assumes for different values of  $r$  terminating the test in acceptance when  $x$  falls below the line. Similarly, the condition  $S_r \leq B$  implies

$$m^x(2-m)^{r-x} \geq \frac{1}{B},$$

$$\therefore x \geq \frac{\log B + r \log (2-m)}{\log (2-m) - \log m} \quad \dots \quad (vi)$$



The example on p. 855 will suffice to illustrate the use of (v) and (vi), since we have made  $p_a = \frac{1}{2}$  and  $p_b = \frac{2}{3} = \frac{4}{3}p_a$  so that  $m = \frac{4}{3}$ . For the 8-fold sample we there cited  $x \leq 1$  for  $S_r \geq 9$  and  $x > 7$  for  $S_r \leq 0.1$ . Thus (v) and (vi) become

$$\begin{aligned} x &\leq \frac{\log 9 + 8 \log 2 - 8 \log 3}{\log 2 - \log 4}; \\ x &\geq \frac{8 \log 2 - 8 \log 3 - \log 9}{\log 2 - \log 4}. \end{aligned}$$

This gives as the lower value  $x \leq (0.4450 \div 0.3011) \simeq 1.4$  so that only scores of  $x = 0$  or 1 satisfy the acceptance criterion and  $x \geq (2.3726 \div 0.3011) \simeq 7.8$  so that only the score  $x = 8$  satisfies the rejection criterion as we have already shown directly.

If we invoke a sequential test of this type, we accept the onus of assigning to the parameter  $m$  a meaningful numerical value. In many biological enquiries, such as the clinical trial, we have little to suggest an appropriate figure. To get this dilemma into focus, let us recall the urn model at the beginning of this section. We shall denote by  $p_1$  the proportion of red balls in Urn I and by  $p_2$  the proportion of red balls in Urn II. The complete set-up is then as follows :

<i>Concordances</i>	$(+ +)$	$p_1 p_2$	$(- -)$	$(1 - p_1)(1 - p_2);$
<i>Discordances</i>	$(+ -)$	$p_1(1 - p_2)$	$(- +)$	$(1 - p_1)p_2.$

Our null hypothesis is that

$$\frac{p_1(1-p_2)}{p_1(1-p_2)+p_2(1-p_1)} = \frac{1}{2} = \frac{p_2(1-p_1)}{p_1(1-p_2)+p_2(1-p_1)}.$$

This is necessarily true if Urn II contains the same proportion of red balls as does Urn I, i.e.  $p_1 = p_2$ . The alternative hypothesis implies that

$$\begin{aligned} \frac{m}{2} &= \frac{p_2(1-p_1)}{p_2(1-p_1) + p_1(1-p_2)} = \frac{p_2 - p_1 p_2}{p_2 - 2p_1 p_2 + p_1}, \\ \therefore m(p_2 - 2p_1 p_2 + p_1) &= 2p_2 - 2p_1 p_2; \\ mp_1 &= (2-m)p_2 + 2p_1(m-1)p_2; \\ p_2 &= \frac{mp_1}{2-m+2p_1(m-1)}. \end{aligned}$$

If we can cite a reliable figure for  $p_1$  referable to the first urn, we can now interpret  $m$  in terms of operational advantage, i.e.  $p_2 - p_1$ . Otherwise, the outcome of the test has little bearing on its presumptive aim.

## 20.06 ESTIMATION AND CONFIDENCE

The need to distinguish between two types of statistical inference as *conditional* and *unconditional* arises from two domains in which we can apply statistical principles in everyday life, namely . (a) the regulation of affairs ; (b) the process of discovery. In commerce, manufacture and social administration we need rules of conduct to safeguard us against known risks. Our concern is then less with the truth of a principle than with the consequences to ourselves when it happens to be true or false, as the case may be. We do not therefore invoke statistical reasoning to justify our confidence in the truth of a particular hypothesis. All we need to know is how often a particular course of action will expose us to the penalties of rejecting its truth. This essentially conditional conception of the use of statistics has a long history in connexion with



the rise of insurance, and is unexceptionable in its right place ; but it has little relevance to what men of science have hitherto regarded as the proper goal of scientific research. If the legitimate scope of statistical reasoning has no concern with impersonal judgments about the truth of hypotheses, we must either dismiss its commonly asserted claim to be an essential component of scientific interpretation or abandon the traditional ethic of the scientific worker.

Our examination of the use of test procedures in 20.03–20.04 has led us to the following conclusions :

- (i) the issues raised in the construction of the balance sheet of Thomas Bayes refer to situations in which the end in view is to pass judgment on the truth of one or other of a set of hypotheses referable to existent populations at risk ;
- (ii) since the terms of reference of conditional statistical inference, as here defined, exclude judgments of this sort, the balance sheet of Bayes has no bearing on the admitted usefulness of decision tests within the legitimate province of conditional assertions ;
- (iii) if circumstances entitle us to limit the range of admissible hypotheses which invite a verdict, it may be possible to devise decision tests which set an acceptable limit on the probability of a false choice without assuming that more than one such hypothesis is indeed referable to an existent population at risk ;
- (iv) since such circumstances will rarely arise in the conduct of research, the scope of unconditional statistical inference is very restricted, if exclusively reliant on test procedure ;
- (v) if we hold that unconditional statistical inference is an important feature of scientific interpretation, we can therefore justify its claims as such by invoking an alternative procedure.

In statistical theory the term *interval estimation* has lately acquired a special meaning to denote such an alternative. We can define it naïvely as in 20.00 to emphasise what distinguishes the end in view from that of the decision test, or in a more sophisticated way, as later, to exhibit the pattern common to each procedure. At the outset, however, it is essential to be clear about what we do *not* mean by estimation in this context. Emphatically, we do not here mean, as is implicit in the all too common expression *the best estimate*, a procedure which assigns some unique value to a parameter of a universe, e.g. the proportion of cards of a particular denomination in a pack or the true mean score of the toss of a die.

The use of the epithet *best* (and the preceding definite article) qualifying the word estimate is so widely current as to justify a brief digression, since its literal semantic implications are indeed wholly inconsistent with a modern attitude to the sort of statements which estimation procedure undertakes to justify. In short, we approach the problem of estimation with an emotional block at the outset, unless we realise that the end in view is not to cite some single best figure. A so-called *point-estimate* is not one to which we can assign an acceptable uncertainty safeguard within the domain of unconditional statistical inference. In different contexts, writers on statistics attach the adjective *best* to a unique sample estimate to signify that it is : (a) unbiased ; (b) most efficient ; (c) sufficient. The first merely means that we arrive at it by a process which would lead us to the right answer if we repeated it on similar samples sufficiently often, an assertion which does not get us very far if we have only one sample at our disposal. The second signifies that the method we use to derive it would ensure the least uncertainty about the range in which it lies ; but does not imply a precise specification of the location of the latter. The third signifies that it embodies all the relevant information the sample supplies ; but tells us nothing about what conclusions the relevant information entitles us to draw.



With these distinctions in mind let us consider a conundrum beloved by pedagogues of the more prosy sort. If I have tossed ten times a penny which came down heads upwards every time, what can I conclude about the probability that it will come down heads next time? If we denote the single spin probability of heads by  $p$ , the best estimate of  $p$  in each sense of the term as used above is then  $p = 1$ , since (i) the observed proportionate sample score of a binomial variate is an unbiased estimate of the unique parameter  $p$ ; (ii) the variance of the binomial distribution is zero when  $p = 1$ ; (iii) the parameter  $p$  suffices to define the distribution for an assigned sample size. My so-called best estimate will thus signify that the penny bears a head on both faces.\*

The modern viewpoint associated with the terms confidence and *interval estimation* repudiates the undertaking to make any such statement. What it does undertake is to specify a range of values (*confidence interval*) within which a parameter lies. In conformity with our initial definition of statistical inference (20.00) such a specification presupposes an *uncertainty safeguard*. In other words, we say in effect: (a) the appropriate answer (i.e. limits assigned to the range) will not be invariably true; but it will be right nine times out of ten or ninety-nine times out of a hundred or with whatever corresponding frequency you care to assign; (b) how precise an answer I can give you (i.e. how narrow the prescribed limits of the range) will depend on how much liability to error you are willing to condone; (c) when you have made up your own mind about how much fallibility you will concede to me in return for how much I may legitimately claim for a more definite assertion, we can get down to business. In terms of what we now call *confidence intervals* the argument proceeds as a public symposium in which the possibility of concord is contingent on an accepted framework of precisely specified fallibility.

The development of the theory is largely due to J. Neyman. In retrospect, and for reasons given in 16.05, Gosset's pioneer paper on the  $t$ -distribution has a special interest in connexion with its beginnings, because it opened the door to an exact method of estimation of the mean of a normal universe; and it is reasonable to suppose that a premonition of its relevance from that viewpoint motivated Fisher's appreciation of its importance before it gained recognition. Credit for what is seemingly the earliest explicit statement of the common sense of the *confidence* approach in the taxonomic domain is due to Wilson (1927). Since Wilson's contribution has received very little recognition, it will not be out of place to quote his words:

In 1927 I called attention to the fact that many statements about probability are highly elliptical and illustrated the matter by the simple case of a point-binomial universe with unknown probability  $p$  and observed value  $p_o$  in some sample. Using the admittedly rough estimate of probability based on the standard deviation one ordinarily writes

$$p_o - \lambda\sqrt{p_o q_o/n} < p < p_o + \lambda\sqrt{p_o q_o/n}$$

and states that the probability that the true value  $p$  in the universe lies between the limits given may be had from a probability-integral table entered with a normal deviation of  $\lambda$  units. I urged

\* Kendall's (1946) remarks (*Advanced Theory of Statistics*, Vol. II, p. 2) are eminently quotable in this context:

"It will clarify our ideas considerably if we draw a distinction between the method or rule of estimation, which, following Pitman, we shall call an Estimator, and the value to which it gives rise in particular cases, the Estimate. The distinction is the same as that between a function  $f(x)$ , regarded as defined for a range of the variable  $x$ , and the particular value which the function assumes, say  $f(a)$ , for a specified value of  $x$  equal to  $a$ . Our problem is not to find estimates, but to find Estimators. We do not reject a method because it gives a bad result in a particular case (in the sense that the estimate differs materially from the true value). We should only reject it if it gave bad results in the long run, that is to say, if the population of possible values of the estimator were seriously discrepant with the value of  $\theta$ . The merit of the estimator is judged by the population of estimates to which it gives rise. It is itself a random variable and has a distribution to which we shall frequently have occasion to refer."

The reader will note that the best *estimator* so defined is not a number, but a rule which would lead us to find a number in the most economical and reliable way, if consistently followed.



that a better procedure *would* be to use for the standard deviation the value  $pq/n$  obtained from the unknown  $p$  of the universe which leads to

$$\frac{p_0 + t/2}{1 + t} - \frac{\sqrt{p_0 q_0 t + t^2/4}}{1 + t} < p < \frac{p_0 + t/2}{1 + t} + \frac{\sqrt{p_0 q_0 t + t^2/4}}{1 + t}.$$

(*Proc. Nat. Acad. Sci.*, Vol. 42, 1942)

We have touched on the elements of confidence theory in Chapter 5 of Vol. I; but it will not be redundant to retrace our steps, if only to emphasise a nicety of definition. In conformity with the usage of the founding fathers of the theory of probability, many statisticians prefer to restrict the term probability in a practical sense to denote the long-run frequency of an external occurrence in contradistinction to the long-run frequency of a correct judgment concerning such an occurrence. If so confidence is not probability. From a behaviouristic viewpoint, however, there is no obvious objection to the use of the same term for the frequency of events which respectively do or do not involve human behaviour at the verbal level, if we make the distinction explicit when appropriate, recognising in what situations it is so. When we take a sample from a *homogeneous* universe, the parameter ( $p$ ) concerning which we seek an estimate has a unique value, the probability of which in any meaningful sense is unity. Thus we can speak with propriety of  $p$  as the probability that it actually lies within those limits if, and only if  $p$  has one of two values, *viz.* :  $p = 0$  or  $p = 1$ ; but we may state with propriety, and without any such restriction, that  $p$  is the probability of a correct assertion about limits between which it lies. The probability of making a correct assertion by consistent application of a rule, and that is what we here mean by *confidence*, may indeed have any value between 0 and 1 in conformity with the way in which we define the limits our assertion sets to the values  $p$  may take.

It will be easy to get into focus the formal implications of this distinction, if we examine artificial situations for which we can construct a model game-of-chance. Those which follow in this section prescribe a homogeneous universe concerning which we seek to estimate some single definitive parameter.

*Model Ia.* We shall conceive that a lottery wheel has 1024 sectors labelled with scores  $x, (x + 1), (x + 2), (x + 3) \dots (x + 9), (x + 10)$  respectively allocated to 1, 10, 45, 120, 210, 252, 210, 120, 45, 10, 1 sectors. So much we know; but we do not know the numerical value of  $x$  itself. At each spin we record as our score that of the sector opposite a fixed pointer. We now suppose that we spin the wheel 40 times and record the mean score ( $M_x$ ) of the 40-fold sample as 6.3. Our problem is to define what we can legitimately assert about  $x$ .

The long-run mean value ( $M$ ) of the score of any sample is, of course,  $(x + 5)$ ; and the terms of  $(\frac{1}{2} + \frac{1}{2})^{10}$  define the u.s.d. of the universe with variance  $\sigma^2 = 2.5$ , whence that of the distribution of the 40-fold sample mean is

$$\sigma_m^2 = \frac{\sigma^2}{40} = \frac{1}{16}.$$

Thus  $\sigma_m = 0.25$ ; and the error involved in a normal quadrature for the distribution of the sample means is trivial. We can thus say that

(a) the mean ( $M_x$ ) of 2.5 per cent. of all samples in the long run will exceed

$$M + 2\sigma_m = M + 0.5;$$

(b) the mean of 2.5 per cent. of all samples in the long run will be less than

$$M - 2\sigma_m = M - 0.5;$$

(c) the mean of 95 per cent. of all samples will lie in the range  $M \pm 2\sigma_m = M \pm 0.5$ .



To say of 2.5 per cent. of all samples that  $M_x > M + 2\sigma_m$  is to say of 2.5 per cent. of all samples that  $M_x - 2\sigma_m > M$ , in which event  $M < M_x - 0.5$ . To say of 2.5 per cent. of all samples that  $M_x < M - 2\sigma_m$  is to say of 2.5 per cent. of all samples that  $M_x + 2\sigma_m < M$ , in which event  $M > M_x + 2\sigma_m$ . The assertion that the sample value ( $M_x$ ) lies within the range  $M \pm 2\sigma_m$  will be true of 95 per cent. of all samples, i.e. the probability that it is true is 0.95. Since such an assertion is formally identical with the statement that  $M$  itself lies within the range  $M_x \pm 2\sigma_m$ , we can assign a probability of 0.95 (95 per cent. confidence level) to the truth of the assertion that  $M$  lies within the range  $6.3 \pm 0.5$ , i.e. from 5.8 to 6.8 inclusive. Since  $M = x + 5$  by definition, we can say with equal confidence that  $x$  itself lies within the range 0.8 to 1.8 inclusive, assigning 1 as the correct value (at the 95 per cent. confidence level), if  $x$  is an integer.

## A CONFIDENCE MODEL

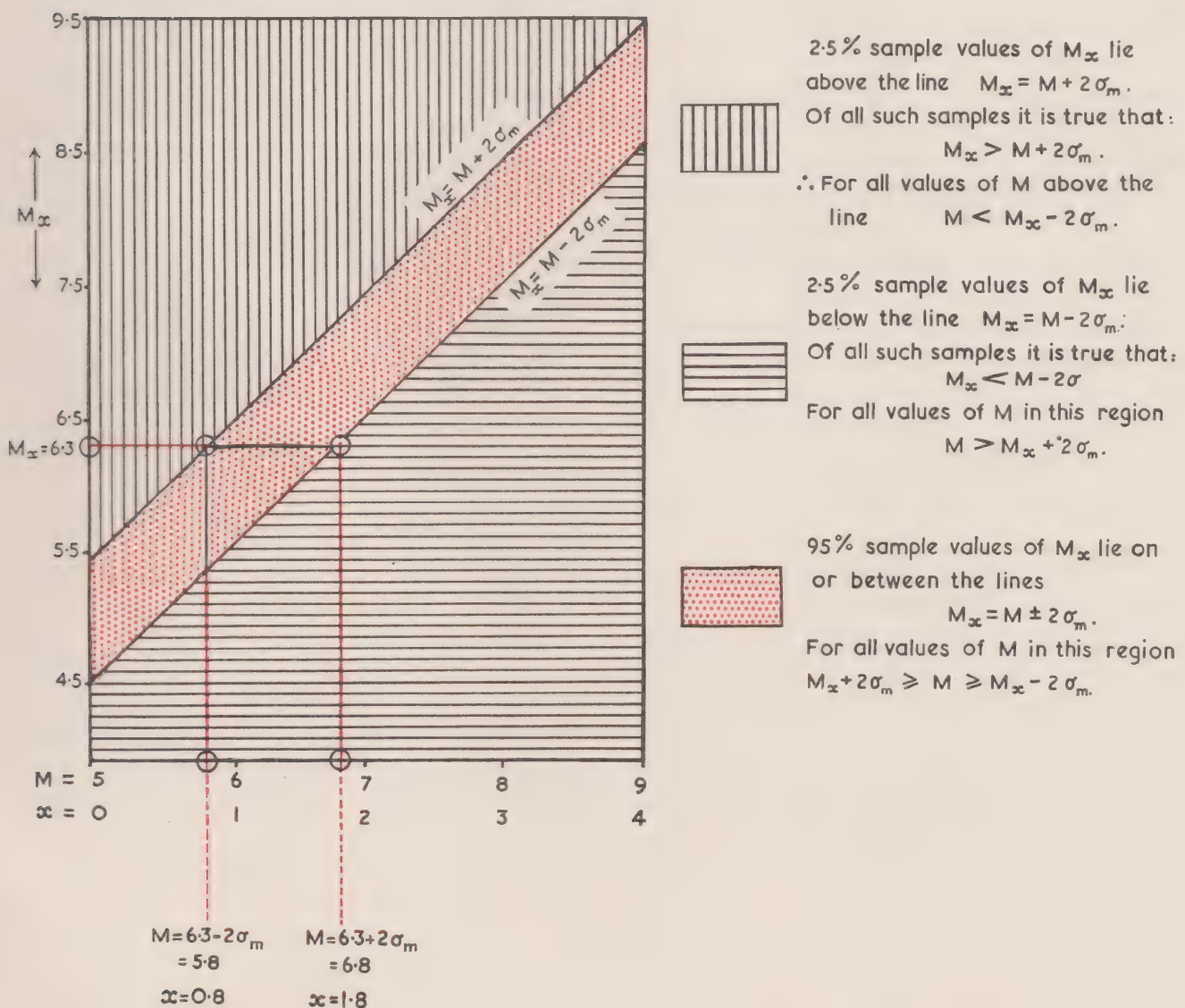


FIG. 137. The Region of Confidence for the lottery wheel model of p. 888.

We can set out the above reasoning in tabular form thus :

<i>Event</i>	<i>Probability of its occurrence</i>	<i>Equivalent assertion</i>	<i>Probability of its truth</i>
$M_x > M + 2\sigma_m$	0.025	$M < M_x - 2\sigma_m$	0.025
$M_x < M - 2\sigma_m$	0.025	$M > M_x + 2\sigma_m$	0.025
$M - 2\sigma_m \leq M_x \leq M + 2\sigma_m$	0.95	$M_x + 2\sigma_m \geq M \geq M_x - 2\sigma_m$	0.95

Fig. 137 exhibits the argument based on our lottery wheel model within the range of values  $5 \leq M \leq 9$  and  $0 \leq x \leq 4$ . For any value of  $M$  we deny the occurrence of all values of  $M_x$  greater than  $M + 2\sigma_m$  or less than  $M - 2\sigma_m$  with a probability of erroneous assertion approximately equal to 0.05. Thus 95 per cent. of all sample values of  $M_x$  will lie within the two *parallel* lines  $M_x = M + 2\sigma_m$  and  $M_x = M - 2\sigma_m$ . There will correspond to any observed value of  $M_x$  two values of  $M$  where the line through  $M_x$  parallel to the abscissa cuts these two lines. These two values will define the range of  $M$  consistent with the probability of error assigned to our denial of the limits of admissible values of  $M_x$ .

In one respect, the foregoing model is highly artificial, i.e. we know in advance the numerical value of the variance ( $\sigma$ ) of the u.s.d. and hence that of  $\sigma_m$ . When sampling from a putatively normal universe we rarely, if ever, have such knowledge ; but we do know the distribution of the ratio  $(M_x - M)$  to the unbiased sample estimate ( $s_m$ ) of  $\sigma_m$ . Hence as explained in 16.09 we can get from the  $t$ -table upper and lower limits for  $M$  consistent with any assigned probability of erroneous statement within the framework of repeated application of the rule. We can do the same for  $\sigma^2$  itself by recourse to tables of the Chi-Square distribution as explained in 16.03.

*Model Ib.* We now suppose that our lottery wheel has 100 sectors on each of which the number of pips is either 0 or 1. We do not know the number ( $100q$ ) of sectors which carry no pips or of sectors  $100p = 100(1 - q)$  which carry one pip. We spin it 100 times and record the mean score of the 100-fold spin as 0.62. Our problem is to define confidence limits of  $p$ , the proportion of sectors which carry one pip. We are here sampling in an infinite two-class universe, and successive terms of  $(q + p)^{100}$  define the frequencies of the observed proportionate (mean) score  $p_o = 0, 0.01, 0.02, 0.03, \dots 0.99, 1.0$ . The *unknown* variance of the distribution of  $p_o$  is given by

$$\sigma_p^2 = \frac{p(1-p)}{100}.$$

Throughout the range of prescribed values other than  $p = 0$  or  $p = 1$  the distribution of the observed proportionate score will be approximately normal. The range  $p_o = p \pm 2\sigma_p$  will therefore define the 95 per cent. confidence level well enough for heuristic purposes. Since  $\sigma_p$  depends on  $p$ , being zero when  $p = 0$  or  $p = 1$ , the two boundaries of acceptable values of  $p_o$  will not be parallel straight lines as in Fig. 137. They will meet at  $p = 0$  and  $p = 1$ , the upper being concave downwards, the lower being concave upwards. The corresponding acceptable range of  $p$  values for any observed value of  $p_o$  is obtainable graphically (Fig. 138), as before, by drawing a horizontal line to parallel the abscissa ; but each limit is subsumed by the two roots of the quadratic

$$(p_o - p)^2 = 4\sigma_p^2 = \frac{pq}{25}.$$

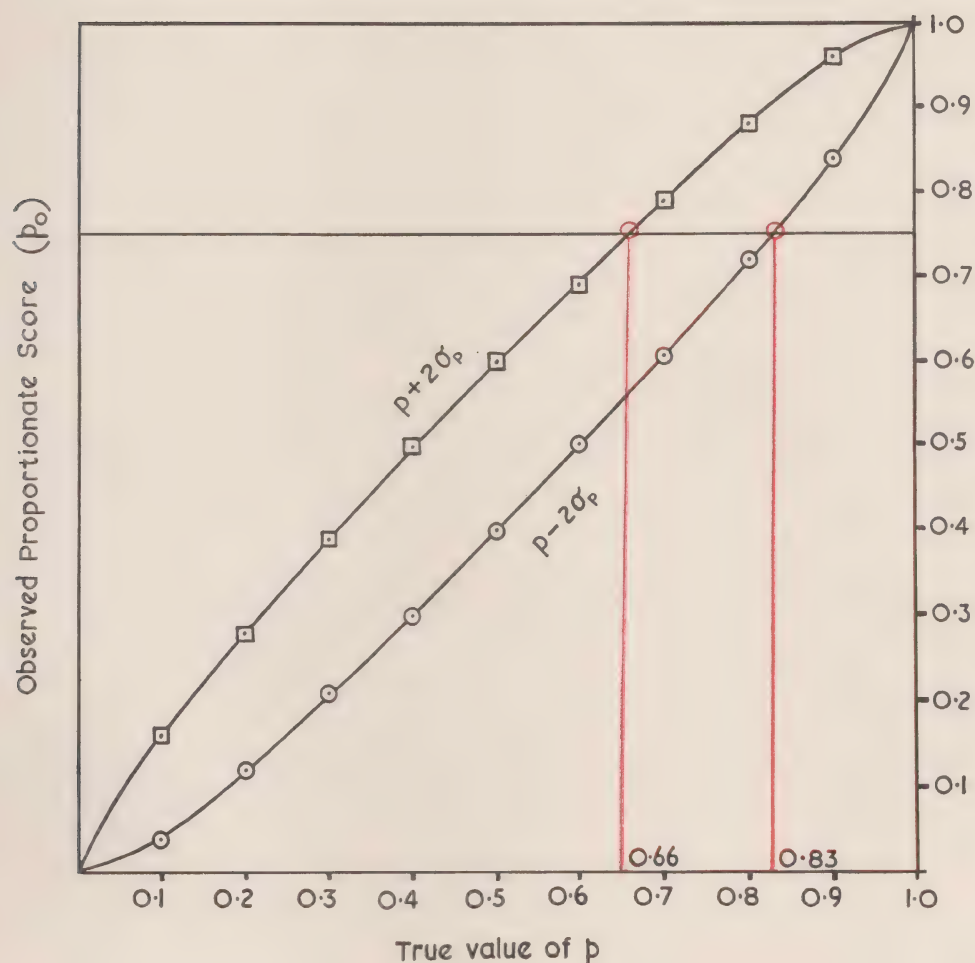
If the observed mean value is 0.62 this becomes

$$\begin{aligned} 25(0.62 - p)^2 &= p(1 - p), \\ \therefore 26p^2 - 32p + 9.61 &= 0, \\ \therefore p &\simeq 0.52 \quad \text{or} \quad 0.71. \end{aligned}$$



CONFIDENCE IN THE DOMAIN OF THE  
2-CLASS UNIVERSE

$$\sigma_p^2 = \frac{pq}{r}; r=100.$$



$$p = \frac{(rp_0 + 2) \pm \sqrt{(rp_0 + 2)^2 - r(r+4)p_0^2}}{r+4}$$

$$\approx 0.66 \text{ or } 0.83$$

FIG. 138. Graphical Interpretation of the Quadratic Solution of Confidence Limits in the Taxonomic domain.

At the  $2\sigma$  confidence level we shall therefore say that our lottery wheel has no more than 71 and no less than 52 sectors carrying one pip. There is no need to generalise this case, already dealt with in Chapter 5 of Vol. I. Four points call for comment:

- (i) the foregoing method is valid only if we can assume that the normal curves give a good quadrature for the binomial distribution, the relevant condition being as stated in 14.07;
- (ii) even so, we can assign the appropriate confidence level correctly only if we pay attention to the half interval correction;

- (iii) if the size of the sample is small we can define, for any value of  $p$ , limits which exclude at least 2.5 per cent. (or other agreed figure) at either end of the range by recourse to the tables of the binomial ; \*
- (iv) if the size of the sample is large, a very small observed value of  $p_o$  may suggest that the Poisson distribution gives a better summation of terms than the normal integral.

*Model Ic.* It limits our horizon unduly, if we confine our interpretation of confidence limits to situations in which we can assume without appreciable error that our score distribution is approximately normal and the confidence interval itself expressible in terms of its variance. The latter has no relevance when the universe is rectangular ; and we may therefore deepen our insight into the logic of confidence theory, if we now lay aside any preoccupations with the normal distribution. As an elementary example of the confidence approach to estimation in the rectangular universe, we may consider the following model. A lottery wheel has  $s$  sectors with consecutive scores from 1 to  $s$ , so that the proportion of sectors whose score value ( $x$ ) exceeds  $m$  ( $\leq s$ ) is  $(s - m) \div s$ . We shall suppose that we spin it once and record  $x$ . Our first problem will be : what can we legitimately say about  $s$  ?

In the treatment of the foregoing model we have side-stepped a limitation of interval estimation in the domain of discrete score values by assuming a good enough normal fit. Unless we postulate a continuous distribution we cannot in fact assign an uncertainty safeguard ( $P_f = \alpha$ ) or confidence level  $(1 - P_f) = (1 - \alpha)$  to an admissible range of score values. The best we can assert is a statement of the form  $P_f \leq \alpha$  or  $P_f < \alpha$ , as when we use tables of the binomial in the situation of Model I (b). One reason for this is that we can assign more than one value to  $m$  consistent with a fixed value  $P_f = \alpha$  for the rule to disregard all samples if  $x > m$ . If the score  $x$  is an integer, e.g.  $k$  or  $(k + 1)$ , we can postulate an infinitude of values to which we can assign the probability  $\alpha$  that  $x > m$  in the range  $k < m < k + 1$ .

It will be convenient to write  $P(x > k)$  for the probability that  $x$  exceeds  $k$  and  $P(x \geq k) = P(x > k - 1)$  for the probability that  $x$  is not less than  $k$ . If  $k$  is an integer there are  $(s - k)$  score values in the range  $x > k$  and  $(s - k + 1)$  in the range  $x \geq k$ , whence

$$P(x > k) = \frac{s - k}{s} \quad \text{and} \quad P(x \geq k) = \frac{s - k + 1}{s} \quad . \quad . \quad . \quad (i)$$

If  $k + 1 > m \geq k$  so that  $m$  is either an improper fraction in the interval between  $k$  and  $(k + 1)$  or is the integer  $k$  itself, we may write  $m = (k + \epsilon)$  and  $k = (m - \epsilon)$  for values of  $\epsilon$  in the range  $0 \leq \epsilon < 1$ . When  $\epsilon = 0$ , we may write

$$P(x > m) = \frac{s - m}{s} \quad \text{and} \quad P(x \geq m) = \frac{s - m + 1}{s}.$$

When  $\epsilon > 0$

$$P(x > m) = P(x > k) = \frac{s - k}{s} = \frac{s - m + \epsilon}{s};$$

$$P(x \geq m) = P(x \geq k) = \frac{s - k + 1}{s} = \frac{s - m + 1 + \epsilon}{s},$$

$$\therefore P(x > m) > \frac{s - m}{s} \quad \text{and} \quad P(x \geq m) > \frac{s - m + 1}{s} \quad . \quad . \quad . \quad (ii)$$

\* Tables of the Binomial Probability Distribution, 1950. National Bureau of Standards, Applied Mathematics Series, 6, Washington ; Clopper and Pearson (1934), *Biometrika*, Vol. 26.



We may subsume both equations (i) and (ii) to cover the possibility that  $m$  may or may not be a whole number in the expressions

$$P(x > m) \geq \frac{s - m}{s} \quad \text{and} \quad P(x \geq m) \geq \frac{s - m + 1}{s}$$

Let us now set  $m = \alpha s$ , so that

$$\text{Rule (i): } P(x > \alpha s) \geq 1 - \alpha;$$

$$\text{Rule (ii): } P(x \geq \alpha s) \geq (1 - \alpha) + 1/s > (1 - \alpha).$$

The proportion of all samples whose score  $x$  exceeds  $\alpha s$  is thus *no less than*  $100(1 - \alpha)$  per cent.; and the proportion of all samples whose score  $x$  is not less than  $\alpha s$  is *greater than*  $100(1 - \alpha)$  per cent. We may set out the implications of the foregoing statements as below:

<i>Event</i>	<i>Probability of its occurrence</i>	<i>Equivalent assertion</i>	<i>Probability of its truth</i>	<i>Probability of its falsehood</i>
$x > \alpha s$	$\geq (1 - \alpha)$	$s < \frac{x}{\alpha}$	$P_t \geq (1 - \alpha)$	$P_f \leq \alpha$
$x \geq \alpha s$	$> (1 - \alpha)$	$s \leq \frac{x}{\alpha}$	$P_t > (1 - \alpha)$	$P_f < \alpha$

We may express this by saying that our uncertainty safeguard for the assertion that  $s$  is less than  $20x$  does not exceed 5 per cent. and our uncertainty safeguard for the assertion that  $s$  is at least  $20x$  is less than 5 per cent. On the basis of observations of a single spin with scores respectively  $x = 5$  and  $x = 10$ , our assertions would thus take the following form, if we deem  $P_f \leq \alpha$  as an acceptable level of uncertainty:

	$x = 5$	$x = 10$	$P_f$
Rule (i)	$s < 100$	$s < 200$	$\leq 0.05$
Rule (ii)	$s < 101$	$s < 201$	$< 0.05$

To say that  $s < 100$  in this context is to say that the upper confidence limit is 99. In terms of confidence limits we therefore write the above as

<i>Upper Confidence Limit of s</i>			
	$x = 5$	$x = 10$	$P_f$
Rule (i)	99	199	$\leq 0.05$
Rule (ii)	100	200	$< 0.05$

Why we cannot express our confidence level in the form of an exact specification of the uncertainty safeguard of the form  $P_f = \alpha$  will be clear if we state the foregoing rules in another way. In effect, rule (i) signifies that we propose to disregard all samples if  $x \leq \alpha s$  and rule (ii) that we shall consistently disregard samples if  $x < \alpha s$ . We can get a backstage view of their implications, if we determine the proportion of excluded samples, i.e. the true uncertainty safeguard prescribed by each rule for values of  $s$  in the neighbourhood of 200, when  $\alpha = 0.05$  defines the upper limit of acceptability for our uncertainty safeguard and the sample score is  $x = 10$ . For  $s = 199, 200$  and  $201$  respectively  $\alpha s = 9.95, 10$  and  $10.05$ . By rule (i) we disregard samples whose scores are 9, 10 and 10. The exact probabilities ( $P_f$ ) of doing so are respectively

$$\frac{9}{199}; \quad \frac{10}{200}; \quad \frac{10}{201}.$$

By rule (ii) we disregard samples whose scores are 9, 9 and 10 with probabilities

$$\frac{9}{199} ; \quad \frac{9}{200} ; \quad \frac{10}{201}.$$

Thus the values of  $P_f$  for  $s$  in the neighbourhood of 200 are

$s$	Rule (i)	Rule (ii)
199	$\simeq 0.045$	$\simeq 0.045$
200	$= 0.05$	$\simeq 0.045$
201	$\simeq 0.0497$	$\simeq 0.0497$

Rule (i) will make  $P_f = 0.05 = \alpha$  when  $s$  is an exact multiple of  $20 = \alpha^{-1}$ ; but otherwise  $P_f < \alpha$ . Rule (ii) makes  $P_f$  nearly equal to  $\alpha$  when  $s$  is an exact multiple of 20 but always less than  $\alpha$ .

The fatal fascination of the continuum to the theoreticians of statistics arises from the circumstance that such inequalities do not trouble us; and it is instructive to examine the consequences of invoking the continuous rectangular variate in this situation. To define a rectangular distribution as a continuous variate we have to satisfy two conditions: (a) the probability  $f(x)dx$  that a score lies in the range  $x \pm \frac{1}{2}dx$  is constant for all values of  $x$ , i.e.  $f(x)dx = K \cdot dx$ ; (b) the complete integral is numerically equal to unity, i.e.

$$K \int_1^s dx = 1 = K(s - 1).$$

The probabilities that the score lies in the range from 1 to  $k$  or  $k$  to  $s$  are then exactly

$$P(x \leq k) = \frac{1}{s-1} \int_1^k dx = \frac{k-1}{s-1};$$

$$P(x \geq k) = \frac{1}{s-1} \int_k^s dx = \frac{s-k}{s-1}.$$

We cannot split a range of whole number score values from 1 to  $s$  into two regions  $x \leq k$  and  $x \geq k$  as we can if  $x$  is a continuous variate; but we can often get a good approximation to a sum by integration if we invoke the half interval correction. We then interpret  $x > k$  to mean  $x > (k + \frac{1}{2})$ , so that

$$P(x > k) \simeq \frac{1}{s-1} \int_{k+\frac{1}{2}}^s dx = \frac{s-k-\frac{1}{2}}{s-1},$$

$$\therefore P(x > k) \simeq \frac{2s-2k-1}{2s-2}.$$

To make  $P(x > k) = 1 - \alpha$ , we have

$$\alpha \simeq \frac{2k-1}{2s-2} \quad \text{and} \quad k = \alpha(s-1) + \frac{1}{2}.$$

The statement  $x > \alpha(s-2) + \frac{1}{2}$  is equivalent to

$$s < \frac{x + \alpha - \frac{1}{2}}{\alpha}.$$

If  $\alpha = 0.05$ , this is equivalent to  $s < (20x - 9)$ ; and when  $x = 10$ ,  $s < 191$  with  $P_f \simeq 0.05$ . The value of  $\alpha$  consistent with  $x = 10$  and  $s = 191$  prescribed by rule (i) above is  $\alpha = 10 \div 191 \simeq 0.0523$ ; and the rule states that  $P_f < \alpha$ , i.e.  $P_f < 0.0523$ .



We did not have to face the issue last discussed in the context of Model I (b), because we invoked a normal approximation for the summation of the terms of a truly discrete binomial sample distribution. It is therefore instructive to re-examine the foregoing model situation on the assumption that the score  $x$  is a continuous rectangular variate. We may then interpret  $x \geq k$  as  $x > (k - \frac{1}{2})$  and  $x \leq k$  as  $x < (k + \frac{1}{2})$ . To accommodate all discrete values in the range  $x = 1$  to  $x = s$  inclusive we must accordingly extend the range of the continuous distribution from  $x = \frac{1}{2}$  to  $x = (s + \frac{1}{2})$ . On this understanding, our formal definition of the continuous rectangular distribution has merely to satisfy two conditions: (a) the probability  $f(x)dx$  that a score lies in the range  $x \pm \frac{1}{2}dx$  is constant for all values of  $x$ , i.e.  $f(x)dx = K \cdot dx$ : (b) the complete integral is numerically equal to unity, i.e.

$$K \int_{\frac{1}{2}}^{s+\frac{1}{2}} dx = 1 = Ks \quad \text{and} \quad K = \frac{1}{s}.$$

The probabilities that the score lies in the range from 1 to  $k$  or beyond  $k$  are then expressible as

$$P(x \leq k) = \frac{1}{s} \int_{\frac{1}{2}}^{k+\frac{1}{2}} dx = \frac{k}{s} \quad \text{and} \quad P(x > k) = 1 - \frac{k}{s}.$$

The above statement is exactly true of the discrete distribution, since  $P(x \leq k) = P(x < k + \frac{1}{2})$  if  $x$  is necessarily an integer. In effect, we make our range from  $\frac{1}{2}\Delta x$  to  $s + \frac{1}{2}\Delta x$ , since  $\Delta x = 1$ ; and we may neglect  $\Delta x$  if  $s$  is very large, as we must assume if we invoke the continuous distribution as a descriptive device. We shall then say that the range is from 0 to  $s$ , and admit fractional values of  $x$  consistent with the specification

$$P(x > k) = \frac{1}{s} \int_k^s dx = 1 - \frac{k}{s}.$$

Accordingly, we now proceed on the assumption that  $x$  can have any real value in the range 0 to  $s$ . To make  $P(x > k) = 1 - \alpha$  we then put  $k = s\alpha$ , so that

$$P(x > s\alpha) = 1 - \alpha.$$

Within the framework of the rule implicit in the procedure, we then assign  $(1 - \alpha)$  as the probability of correctly asserting

$$s < \frac{x}{\alpha}.$$

When  $\alpha = 0.05$ , this is equivalent to assigning  $P_f = 0.05$  to the assertion that  $s$  lies within the range from 1 to  $20x$ .

We have hitherto confined our attention to a procedure which entitles us to assign to  $s$  an upper confidence limit with an uncertainty safeguard  $P_f \leq \alpha$ . If we wish to place it with a preassigned uncertainty safeguard in an interval  $ax > s > bx$ , the form of statement we may make is no longer unique. If we may justifiably proceed on the assumption that we can assign an exact uncertainty safeguard  $P_f = \gamma$  to what assertions we do make within the framework of a prescribed rule of procedure, i.e. that we may legitimately rely as above on the continuous distribution, we may write

$$P(k < x \leq m) = \frac{1}{s} \int_k^m dx = \frac{m - k}{s}.$$

If we now write  $k = \beta s$  and  $m = \alpha s$

$$P(\beta s < x \leq \alpha s) = \alpha - \beta.$$

We then assign an uncertainty safeguard  $P_f = 1 - (\alpha - \beta)$  to the assertion

$$\frac{x}{\beta} > s \geq \frac{x}{\alpha}.$$

If  $\beta = 0.025$  and  $\alpha = 0.975$  so that  $P_f = 0.05$  our final statement will thus be

$$40x > s \geq \frac{40x}{39}.$$

Now  $P_f = 0.05$  if  $\beta = 0.01$  and  $\alpha = 0.96$ . We are therefore entitled to assign  $P_f = 0.05$  as the uncertainty safeguard to the alternative assertion

$$100x > s \geq \frac{25x}{24}.$$

When we write  $P(x > s\alpha) = 1 - \alpha$  or  $P(\beta s < x \leq \alpha s) = \alpha - \beta$ , we state the probability of an event, i.e. value of unit score  $x$ , within the framework of the classical theory of probability and the convenient fiction that the distribution is continuous. Our assertion signifies: for the fixed value  $s$  of the relevant parameter,  $P_{x.s}$  is the probability that the unit score will lie in such and such a range. We have refrained from writing the probability we assign to the equivalent assertions in the notation

$$P(s < x\alpha^{-1}) = 1 - \alpha \quad \text{or} \quad P(\beta^{-1}x > s \geq \alpha^{-1}x) = \alpha - \beta$$

lest we should hastily interpret them in terms of inverse probability, i.e. as if we could legitimately say: for the fixed value  $x$  of the unit score,  $P_{s.x}$  is the probability that  $s$  will lie in the specified range. Such a form of words is inconsistent with Neyman's theory. We must interpret a statement in the form  $P(ax > s \geq bx) = \gamma$  as a summary of the long-run result of consistently adopting one and the same rule of conduct regardless of the value (e.g.  $x = 5$ ) the score  $x$  may have in any single trial, including the particular trial to which our specification of the interval estimate is referable. The formal statement of the rule will be adequate only if it explicitly specifies  $x$  as an unknown which may assume any value within its admissible range. We misinterpret it if we condense our verdict in such a form as

$$P\left(200 > s \geq \frac{200}{39}\right) = 0.95.$$

This is an act of self deception into which we easily slide, if we write the formal identities:

$$\beta(h + dh) = x = \alpha h;$$

$$\frac{x}{h} - \frac{x}{h + dh} = \alpha - \beta = \frac{x \cdot dh}{h(h + dh)};$$

$$P(h + dh > s \geq h) = \frac{x \cdot dh}{h^2}.$$

We have now eliminated any reference to  $x$  as a variable in the expression on the left and have obtained on the right what is seemingly the element of a probability distribution and satisfies the fundamental property of the latter, if we fix  $x$  and define the range of  $s$  from  $h = x$  to  $h = \infty$ , so that

$$x \int_x^\infty h^{-2} \cdot dh = 1.$$



This step, which leads to what Fisher calls a fiducial probability distribution, is admissible only if we can legitimately confine our statements to situations in which  $x$  has one and the same value (e.g.  $x = 5$ ). We could then write

$$P(s < k) = x \int_x^k h^{-2} \cdot dh = \frac{k - x}{k}.$$

If  $k = 20x$ , we thus obtain by a somewhat circuitous route a result already derived within the framework of the assumed continuous rectangular distribution, i.e.  $P(s < 20x) = 0.95$ . The numerical consistency of many—though not all—results embodied in Fisher's approach to interval estimation with those to which the theory of confidence limits leads us did indeed at one time blind many statisticians to what we now see to be a radical difference. If we conceive  $x \cdot f(h)dh$  as an element of a probability distribution, we have to regard  $h$  and  $x$  as independent to arrive at a numerical result consistent with confidence theory in the continuous domain; but we can do so only if we then treat  $x$  as a constant in the algebraic manipulation. We thus implicitly fix our interval in terms of a preassigned value of  $x$  to arrive at the specification of a probability dependent thereon; but this is inconsistent with the programme of Neyman's theory which specifies the interval in terms of a preassigned probability independent of the outcome of any single trial and hence of any preassigned value of  $x$ .

We come to a parting of the ways, when we ask questions involving joint distributions, such as that of the so-called *Behrens* test invoked to estimate the difference between the means of different normal universes. In the confidence theory of interval estimation, we must first specify the composite distribution of relevant composite score values referable to all possible values of each variate, e.g. that of the  $r$ -fold sample mean difference  $d_{m.s}$  referable to all possible values the sample means ( $M_{a.s}$  and  $M_{b.s}$ ) may assume. The prescription of the Fisher school derives a composite fiducial distribution from the particular fiducial distributions of  $M_{a.s}$  and  $M_{b.s}$ . Unless the variances  $\sigma_a^2$  and  $\sigma_b^2$  of the parent unit sample distributions are equal, the two procedures lead to different results; and this has provoked a lively controversy conducted with an output of heat disproportionate to the illumination conferred.\*

## 20.07 ESTIMATION AND THE BAYES' DILEMMA

We have hitherto regarded the problem of estimation as that of assigning a probability to the truth of the assertion that some definitive parameter of a *homogeneous* universe lies between specified limits. So stated, the issue sidesteps the disquieting dilemma with which the balance sheet of Bayes confronts us. Bayes' theorem is essentially about a *stratified* (heterogeneous) universe, e.g. a bag in which some pennies are unbiased and one penny (through a defect of minting) has the King's head on both sides. In effect, it says: to know how often I should be right in judging a coin taken from the bag as the one defective coin after getting 10 successive heads in a single 10-fold toss, I must also know how many other coins the bag contains. The dilemma to which the theorem draws attention is that we rarely have such knowledge; but it is one which the theory of confidence sidesteps. The confidence theory of interval estimation deals with a *Bernoullian universe*, e.g. a bag containing pennies *all of the same sort*; and formulates how much we can say with propriety about the behaviour of such pennies on the basis of a single 10-fold sample. As we have seen, it relinquishes the attempt to assign a so-called *best* value for the long-run frequency ( $p$ ) of heads as the result of tossing any one of them in preference to an exact statement in terms such as the following: though I cannot say what is the

\* Behrens, W. V. (1929), *Landw. Jb.*, **68**, 807. Fisher, R. A. (1935), *Ann. Eug. Lond.*, **6**, 391. Sukhatme, P. V. (1938), *Sankhya*, **4**, 39. Neyman, J. (1941), *Biometrika*, **32**, 128.



correct or best value of  $p$ , I can tell you within what limits  $p$  will lie if you will agree to let me be wrong not more than  $aN$  times (e.g.  $0.05N$ ) in  $N$  trials when  $N$  is very large. It is the writer's belief that Neyman (1934) does not overstate the novelty or the importance of the viewpoint explored in 20.06 when he declares :

The solution of the problem which I described as confidence intervals has been sought by the greatest minds since the work of Bayes 150 years ago. Any recent book on the theory of probability includes large sections concerning this problem. The present solution means, I think, not less than a revolution in *the theory* of statistics.—(*J. Roy. Stat. Soc.*, Vol. 97, p. 536.)

We shall now examine model situations which suggests an alternative more sophisticated approach to the problem of confidence to clarify what is common to the domain of decision tests and the domain of estimation. It is also of special interest for another reason mentioned at the end of 20.03 above. Till recently, it has been common to assume that an adequate theory of statistical decision must come to terms with the prior probabilities of Bayes' theorem. This belief leads to an impasse unless we are content to embrace the highly exceptionable postulate mentioned in 20.03 ; but it rests on a debatable assumption that the model situation with which Bayes' theorem deals is factually relevant to statistical decisions involving no more than one hypothesis referable to an existent population at risk. We can best see the irrelevance of the theorem to the issue of estimation if we : (a) provisionally postulate a model situation to which it is indeed factually relevant ; (b) formulate a procedure which is valid for all conceivable values of the prior probabilities and therefore to the limiting case when there is only one urn. The universe of our models of this section will be a stratified universe, and our problem to attach an acceptable uncertainty safeguard to the assertion that a parameter definitive of the single stratum from which we take a particular sample lies within a specified range.

*Model IIa.* With this end in view we shall suppose that someone spins 40 times one of 100 lottery wheels chosen at random. Each such wheel has 1024 sectors like the wheel of our first model in 20.06 with scores of  $x, x + 1, (x + 2), \dots (x + 9), (x + 10)$ , allocated respectively to 1, 10, 45,  $\dots$  10, 1 sectors. The recorded score of the 40-fold spin is again 6.3, and we do not know the value of  $x$  associated with the particular wheel selected for the spin. We do know, however, that each wheel is one of eleven types as follows :

Type	No. of wheels	Value of $x$
I	1	0.5
II	3	0.6
III	10	0.7
IV	17	0.8
V	20	1.1
VI	7	1.3
VII	12	1.5
VIII	3	1.8
IX	8	1.9
X	2	2.0
XI	17	2.1

In this model set-up, we may construct 11 admissible hypotheses about the value of  $x$ , and hence of the expected mean  $M = (x + 5)$ . For each hypothesis the standard deviation of the distribution of the observed mean ( $M_x$ ) of the 40-fold spin is  $\sigma_m = 0.25$ , and to each hypothesis we can assign a prior probability in Bayes' sense. From this point of view, the relevant information is as follows :



<i>Hypothesis</i>	<i>Prior Probability</i>	<i>M</i>	$(M - M_x) \div \sigma_m$
I	0.01	5.5	- 3.2
II	0.03	5.6	- 2.8
III	0.10	5.7	- 2.4
IV	0.17	5.8	- 2.0
V	0.20	6.1	- 0.8
VI	0.07	6.3	0
VII	0.12	6.5	+ 0.8
VIII	0.03	6.8	+ 2.0
IX	0.08	6.9	+ 2.4
X	0.02	7.0	+ 2.8
XI	0.17	7.1	+ 3.2

We shall now make in the following rule. We shall reject some hypotheses as inadmissible and reserve judgment on others which we shall accordingly regard as admissible, applying to each hypothesis the same criterion of rejection, i.e. that it assigns to the deviation of the observed score ( $M_x = 6.3$ ) from the expected value ( $M$ ) prescribed by the particular hypothesis a value *numerically* greater than  $2\sigma_m$ . We then reject all hypotheses except IV–VIII inclusive, and are left with the assertion that  $M$  lies in the range 5.8–6.8 corresponding to values of  $x$  from 0.8 to 1.8.

Our uncertainty safeguard for rejection of every hypothesis when true is 0.05 since our rejection criterion is modular. That the unconditional uncertainty safeguard for the final verdict is also 0.05 as for Model Ia, we may make explicit as follows. We first remind ourselves that we can falsely *reject* only one hypothesis, since only one can be true. Thus the unconditional probability of a false verdict is the unconditional probability of falsely rejecting one or other of an exclusive set of hypotheses, and is therefore obtainable by recourse to the addition rule. If  $P_h$  is the prior probability that the particular hypothesis  $H$  is applicable to the situation, i.e. that we chose at random a wheel of type  $H$  to spin, the probability of falsely rejecting it is  $\alpha P_h$ ; and by definition

$$\sum_{h=1}^{h=11} P_h = 1.$$

The probability of making a false decision is the probability of falsely rejecting any one of the hypotheses, i.e.

$$\sum_{h=1}^{h=11} P_h \cdot \alpha = \alpha \sum_{h=1}^{h=11} P_h = \alpha.$$

Thus  $\alpha$  is our uncertainty safeguard to the assertion that  $M_x$  lies within the prescribed limits; and *the prior probabilities of Bayes do not affect its value*. We arrive at exactly the same result as for the corresponding situation (Model Ia) of 20.06, where we set the same uncertainty safeguard to the same range of admissible values of the parameter  $x$  of *one and the same wheel*.

Should the reader find the last step of this argument difficult to follow, it may be helpful to set it out in the form of a truth table for which it will suffice to predicate only 4 hypotheses with prior probabilities  $P_1, P_2, P_3, P_4$ , and corresponding definitive parameters  $M_1, M_2$ , etc. Each hypothesis corresponds to a fictitious possible universe, which is a wheel of a given type in our model situation. The hypothesis we deem to be applicable to the situation is the actual

universe from which our sample comes, i.e. the particular wheel of which we record the outcome of a 40-fold spin. We may then set out the procedure in stages as follows :

Prior Probability of choosing the wheel	Probability of Rejection as such ( $\alpha$ )	Probability of Retention as possibly such ( $1 - \alpha$ )
$P_1$	$P_1\alpha$	$P_1(1 - \alpha)$
$P_2$	$P_2\alpha$	$P_2(1 - \alpha)$
$P_3$	$P_3\alpha$	$P_3(1 - \alpha)$
$P_4$	$P_4\alpha$	$P_4(1 - \alpha)$
Total	$\alpha(P_1 + P_2 + P_3 + P_4) = \alpha$	$(1 - \alpha)(P_1 + P_2 + P_3 + P_4) = (1 - \alpha)$

In this table each cell entry of the second column records the probability that we shall both choose a particular wheel to spin and reject the conclusion that we have done so. Each cell entry on the right records the probability that we suspend judgment. The grand total of all the cell entries is  $\alpha + (1 - \alpha) = 1$ . Thus our decisions are classifiable exclusively and exhaustively as either *definitely false* or *uncertain*, and we may interpret our balance alternatively in terms of probabilities assignable to our decisions as follows :

Hypothesis	Decision	
	False	Non-committal
I	$P_1\alpha$	$P_1(1 - \alpha)$
II	$P_2\alpha$	$P_2(1 - \alpha)$
III	$P_3\alpha$	$P_3(1 - \alpha)$
IV	$P_4\alpha$	$P_4(1 - \alpha)$
Total	$\alpha$	$1 - \alpha$

In the set-up of this Model, we regard any one of a limitless number of values  $p$  may have as a hypothesis referable to a conceivably, but not necessarily, existent population at risk. We thus interpret the process of estimation as a method of screening an exhaustive set of hypotheses as admissible or otherwise by successively applying to each a test prescribing the same probability of rejection if the hypothesis is indeed true. Our universe of hypotheses so conceived is a stratified universe, in which strata with the same definitive parameter  $P_h$  provisionally constitute an existent population at risk with an assignable finite prior probability in the jargon of Bayes' theorem. Bayes' prior probabilities ( $P_h$ ) are then relevant to the initial formulation of the problem ; but *they do not appear in the solution*. Consequently, we are free to assign to the prior probability of any single hypothesis any value in the range 0 to 1 consistent with the restriction that the sum of all the prior probabilities is unity. Whether there corresponds an existent population to a particular hypothesis in our fictitious stratified universe is therefore immaterial. That a particular hypothesis to which we apply the test corresponds to no existent population merely means that  $P_h = 0$ . To conceive the universe as unstratified is to assign  $P_h = 1$  to one stratum and  $P_h = 0$  to every other one. In this sense, Model I is therefore a limiting case of Model II.



This way of looking at the problem of estimation makes the distinction between the domain of test decision and that of estimation less clear-cut than the alternative ; but we should not lose sight of what remains. If we perform a decision test to arbitrate *simultaneously* on the merits of alternative hypotheses which constitute an exhaustive set our rejection criterion or criteria determines which we accept and which we reject ; and we can never assign the same probability of rejection if true to more than 3 hypotheses on this understanding. If we interpret the procedure of estimation in terms of the model of this section, we can regard it as the performance of a battery of tests, but the score value which defines the criterion of rejection is different for each test and the decision to reject any one hypothesis or group of hypotheses does not prescribe acceptance of any other single hypothesis. We *successively* apply to each a test involving a new value of the score deviation ( $x - M$ ) as the criterion which ensures the same probability of rejection for each hypothesis when true. If we assert that one group of hypotheses constitute an *admissible* in contradistinction to a residual group as an *inadmissible* set, we then do so on the assumption that one of the former is identifiable with the correct one.

*Model II (c).* In the homogeneous universe of Model I, we have seen that we can set an upper limit ( $P_f < \alpha$  or  $P_f \leq \alpha$ ) to the uncertainty safeguard we attach to a confidence boundary in the domain of discrete score values ; but we cannot make an exact statement of the form  $P_f = \alpha$ . Let us now therefore look at the problem raised by Model I (c) of 20.06 as one of sampling in a stratified universe. We shall postulate as below an assemblage of 100 lottery wheels of 12 types with consecutive scores 1 to  $m$  inclusive if  $s = m$  is the number of sectors of a wheel of type  $H$ . Thus we have 12 hypotheses about  $s$  to explore, each referable to an existent population at risk ; and we shall once more limit our decisions to rejection and reservation of judgment. We know the score  $x$  of a single spin without knowing the type of wheel to which it is referable. Our problem will be to assign a probability to an admissible set of hypotheses.

Type of Wheel ( $H$ )	No. of Sectors ( $s_h$ )	No. of Corresponding Wheels ( $N_h$ )	Prior Probability of Choice ( $P_h = N_h \div 100$ )
1	5	13	0.13
2	19	2	0.02
3	20	1	0.01
4	21	3	0.03
5	39	7	0.07
6	40	12	0.12
7	99	3	0.03
8	100	4	0.04
9	101	9	0.09
10	199	10	0.10
11	200	15	0.15
12	201	21	0.21
	Total	100	1.00

For Model I (c) we formulated two rules

$$\text{Rule (i) } s < \frac{x}{\alpha} \text{ with } P_f \leq \alpha ;$$

$$\text{Rule (ii) } s \leq \frac{x}{\alpha} \text{ with } P_f < \alpha .$$

In effect, the first rule states that we reject the hypothesis  $s = s_h$  unless  $x > \alpha s_h$ ; and the second states that we reject the hypothesis  $s = s_h$  unless  $x \geq \alpha s_h$ . Thus our rejection criteria are

Rule (i) Reject if  $x \leq \alpha s$  with  $P_f \leq \alpha$ ;

Rule (ii) Reject if  $x < \alpha s$  with  $P_f < \alpha$ .

As below, we may then draw up a table of verdicts based on each of the foregoing rules for different experiments in which  $x = 5$  and  $x = 10$  respectively. In each case we assume that  $\alpha = 0.05$  is an acceptable level of uncertainty.

Hypothesis ( $h$ )	No. of Sectors ( $s_h$ )	Criterion ( $\alpha s_h = 0.05 s_h$ )	$x = 5$		$x = 10$	
			Verdict by Rule (i)	Verdict by Rule (ii)	Verdict by Rule (i)	Verdict by Rule (ii)
1	5	0.25	Open	Open	Open	Open
2	19	0.95	Open	Open	Open	Open
3	20	1.00	Open	Open	Open	Open
4	21	1.05	Open	Open	Open	Open
5	39	1.95	Open	Open	Open	Open
6	40	2.00	Open	Open	Open	Open
7	99	4.95	Open	Open	Open	Open
8	100	5.00	REJECT	Open	Open	Open
9	101	5.05	REJECT	REJECT	Open	Open
10	199	9.95	REJECT	REJECT	Open	Open
11	200	10.00	REJECT	REJECT	REJECT	Open
12	201	10.05	REJECT	REJECT	REJECT	REJECT

The range of  $s$  values covered by open verdicts thus corresponds precisely with the outcome of our examination of Model I (c) for which the upper confidence limits are 99 and 199 respectively for  $x = 5$  and  $x = 10$  with  $P_f \leq 0.05$  (Rule i) or 100 and 200 respectively for  $x = 5$  and  $x = 10$  with  $P_f < 0.05$  (Rule ii). The meaning of the correspondence is evident if we recall the meaning of the true conditional uncertainty safeguard ( $P_{f,h}$ ) of hypothesis  $H$  in the domain of discrete score values. If our criterion of rejection is  $x \leq \alpha s$ , we exclude only samples whose score value is  $x = \alpha s$  when  $\alpha s$  itself is an integer. Thus  $P_{f,h}$ , the proportion of excluded score values when hypothesis  $H$  is true, is the ratio to  $s$  of the nearest integer not exceeding  $s$  and is always less than or equal to  $\alpha$ . If  $0 \leq \epsilon_h < 1$  we may thus write

$$P_{f,h} = \alpha - \epsilon_h,$$

$$\therefore P_f = \sum_{h=1}^{h=12} P_h \cdot P_{f,h} = \alpha \sum_{h=1}^{h=12} P_h - \sum_{h=1}^{h=12} P_h \cdot \epsilon_h,$$

$$\therefore P_f = \alpha - \sum_{h=1}^{h=12} P_h \cdot \epsilon_h.$$



Since we have chosen the rejection criterion so that  $P_{f.h} \leq \alpha$ , all values of  $\epsilon_h$  must be zero or positive. Rule (ii) asserts that they are all positive, whence we obtain as for Model I (c),

$$P_f < \alpha.$$

In this instance, some values of  $\epsilon_h$  are positive when we apply Rule (i) and others zero. Thus  $P_f < \alpha$  as before; but this is not inconsistent with the assertion  $P \leq \alpha$  being included therein. A generalised Model II situation must take stock of the possibility that  $P_{f.h} = \alpha$  for each wheel as would be true if we knew that the recorded score referred to a wheel of any one of types 3, 6, 8, 11 above. For each of these  $P_{f.h} = 0.05$  and  $\epsilon_h = 0$  as will be seen by citing the values of  $P_{f.h}$  prescribed by our rejection criterion, *viz.*:

$s_h$	$\alpha s_h$	Rule (i)	Rule (ii)
5	0.25	0.0000	0.0000
19	0.95	0.0000	0.0000
20	1.00	0.0500	0.0000
21	1.05	0.0476	0.0476
39	1.95	0.0256	0.0256
40	2.00	0.0500	0.0250
99	4.95	0.0404	0.0404
100	5.00	0.0500	0.0404
101	5.05	0.0495	0.0495
199	9.95	0.0452	0.0452
200	10.00	0.0500	0.0450
201	10.05	0.0497	0.0497

In the treatment of Model I (c) we have already recognised one reason for regarding the concept of fiducial probability as an inadequate basis for a theory of statistical inference in that it restricts the field of discussion to continuous variates. Further consideration of the model situation we have last discussed gives us an opportunity for contrasting two theories of interval estimation from a different viewpoint. Fiducial probability takes its origin from assumptions common to the theory of confidence; but Neyman's development of the latter is inconsistent with Fisher's interpretation of the former, unless there is some sense in which only one admissible preassigned rule of test procedure is appropriate to one and the same situation. Models I (c) and II (c) do indeed refer to a situation in which only one such rule invites our attention as relevant to the end in view; but we have not excluded the possibility that more than one might each have seemingly equal claims to commend it from a purely formal viewpoint. We shall now examine a situation in which this dilemma arises.

Since the issue has special relevance to the concept of fiducial probability, we shall postulate a continuous rectangular distribution over the range  $\frac{1}{2}$  to  $s + \frac{1}{2}$ , and examine what statements we may make when we draw two unit samples with scores  $x_1$  and  $x_2$ . Two, though not the only two, rules which we may formulate will serve our purpose well enough for heuristic purposes. We shall alternatively seek to prescribe an upper confidence limit to  $s$  with an uncertainty safeguard  $\alpha$  by recourse to

- (i) the maximum score  $x_m$  being  $x_m = x_1$  if  $x_1 \geq x_2$  and  $x_m = x_2$  if  $x_2 \geq x_1$ ;
- (ii) the score sum  $x_{12} = x_1 + x_2$ .

The probability that  $x_m \leq m$  is the probability assignable to the joint occurrence that each score lies in the range from  $x = 0$  to  $x = m$  inclusive, i.e.

$$P(x_m \leq m) = \frac{m^2}{s^2} \quad \text{and} \quad P(x_m > m) = 1 - \frac{m^2}{s^2}.$$

We wish our final assertion to take the form  $s < kx$  with a probability  $(1 - \alpha)$  of correct assertion if we consistently follow the test procedure, whence we write  $P(x_m > m) = (1 - \alpha)$ ,

$$\frac{m^2}{s^2} = \alpha \quad \text{and} \quad P(x_m > s\sqrt{\alpha}) = 1 - \alpha.$$

Within the framework of this rule, we then assign  $\alpha$  as the uncertainty safeguard to the assertion

$$s < \frac{x_m}{\sqrt{\alpha}}.$$

If we base our test procedure on  $x_{12}$  defined as above, the reader unfamiliar with the continuous rectangular distribution will find it helpful first to make a simple chessboard diagram of the 2-fold *discrete* score-sum distribution. It is then evident that we may express the probability that  $x_{12}$  lies in the range 2 to  $k$  if  $x = 1$  is the origin of the unit score distribution in two ways :

$$P(x_{12} > k) = \frac{(2s - k)(2s - k + 1)}{2s^2} \quad \text{when} \quad k > s + 1;$$

$$P(x_{12} > k) = 1 - \frac{k(k - 1)}{2s^2} \quad \text{when} \quad k \leq s + 1.$$

For the continuous case we may represent our chessboard geometrically as a rectangle of area  $s^2$  and the region in which all values  $x_{12} \leq k$  lie when  $k \leq s$  as a triangle of area  $\frac{1}{2}k^2$ . Since we wish to associate a probability  $(1 - \alpha)$  near unity to the truth of the assertion  $s < k^{-1} \cdot x$ , our concern will be with the smaller value of  $k$ . For the continuous case we then write

$$P(x_{12} > k) = 1 - \frac{k^2}{2s^2} = 1 - \alpha,$$

$$\therefore P(x_{12} > s\sqrt{2\alpha}) = 1 - \alpha.$$

Our second rule thus assigns  $\alpha$  as the uncertainty safeguard to the assertion

$$s < \frac{x_{12}}{\sqrt{2\alpha}}.$$

We thus have two rules which assign different values to the upper confidence limit of  $s$  at one and the same confidence level  $(1 - \alpha)$ . In the strictly behaviourist formulation of confidence theory by Neyman this involves an inconsistency only if both rules incorporate all the information about the unknown parameter the sample can supply. One rule may be better than another, if it incorporates more information ; but its use may have drawbacks which outweigh its merit on that account. In Fisher's theory of interval estimation no such freedom of choice is admissible. The avowed intention of the concept of fiducial probability is to express the intensity of legitimate conviction referable to a particular sample. If so, only one rule can be right, namely the rule which invokes all the information the sample supplies. Fisher speaks of a statistic, i.e. sample score, which has this property, as the *sufficient* one.

The two statistics  $x_m$  and  $x_{12}$  used in the foregoing situation will serve to illustrate what is and what is not a sufficient statistic in Fisher's sense of the term, if we now consider  $x_1$  and  $x_2$  as unit samples from a *discrete* rectangular universe with a range of scores from 1 to  $s$  inclusive. In deriving a rule on the basis of either we have suppressed any explicit specification of  $x_1$  and  $x_2$ . If our chosen statistic is defective it can be so only for that reason. We shall therefore ask : have we lost anything by withholding such information ? We may answer this by considering



the consequences of confining our attention in a sequence of trials to samples with some pre-assigned value of  $x_m$  or  $x_{12}$ .

Let us first suppose that the preassigned value of  $x_m = 3$ . The different sorts of double samples that are consistent with this value occur with equal frequency and are specifiable as follows: (1, 3); (2, 3); (3, 3); (3, 2); (3, 1). This set of equally frequent values is the same for all values of  $s$  consistent with the specification  $x_m = 3$ . Thus we have suppressed no information about  $s$  by scoring our sample in this way. Is the same true of  $x_{12}$ ? Let us now consider samples w.r.t. which  $x_{12} = 8$ . This specification is consistent with any value  $s \geq 4$ , but this condition does not suffice to specify what individual values  $x_1$  and  $x_2$  have. If  $s = 4$  the only double sample consistent with the specification  $x_{12} = 8$  is (4, 4). If  $s = 5$ , three paired score values are allowable: (3, 5) (4, 4) (5, 3). If  $s = 6$  we may have: (2, 6) (3, 5) (4, 4) (5, 3) (6, 2). Thus we can say more about  $s$ , if we know the individual score of  $x_1$  and  $x_2$  than we can if we know only the value of the *insufficient* statistic  $x_{12}$ ; but the individual values of  $x_1$  and  $x_2$  tell us no more than we already know, if told the value of the *sufficient* statistic  $x_m$ .

We have now to state the definition of a sufficient statistic formally. To do so we first remind ourselves that to each 2-fold sample specified in terms of the sequence of unit samples we may assign as above a bivariate score, e.g. (3, 5) or (5, 3). We may then speak of  $P_{12 \cdot s}$  as the unconditional probability that any sample has the bivariate score  $(x_1, x_2)$  and  $P_{12 \cdot m}$  as the conditional probability that it has this score if  $x_m$  is the maximum score. In the same sense, we may label the unconditional probability of a multivariate score  $(x_1, x_2, x_3, \dots, x_r)$  definitive of an  $r$ -fold sample as  $P_{(1 \cdot 2 \cdot 3 \dots r) \cdot p}$  for a distribution whose definitive parameter is  $p$  and  $P_{(1 \cdot 2 \cdot 3 \dots r) \cdot x}$  as its conditional probability when the sample statistic is  $x$ , if we can define it from our knowledge of  $x$  alone. We may then define by  $P_{x \cdot p}$  the probability that the sample statistic will be  $x$  if the parameter is  $p$  and obtain by recourse to the product rule

$$P_{(1 \cdot 2 \dots r) \cdot p} = P_{x \cdot p} \cdot P_{(1 \cdot 2 \dots r) \cdot x}.$$

We have now split the unconditional probability of getting the bivariate score which summarises all the information the sample supplies into two factors one of which is independent of  $p$  if the statistic is sufficient, i.e. if (as is true of  $P_{12 \cdot m}$ ) we can specify it without knowing the value of the universe parameter. We thus take as our formal criterion of a sufficient statistic the resolution of the probability of the multivariate score into two factors of which one does not contain  $p$ .

By recourse to a simple chessboard lay-out of  $s^2$  cells with border scores from  $x = 1$  to  $x = s$  inclusive and  $x = 1$  to  $x = s$  inclusive we may amplify this breakdown w.r.t.  $x_m$  and  $x_{12}$  for the discrete rectangular universe. Each cell of the grid is referable to a unique pair of values  $x = x_1$  and  $x = x_2$ , but the same value of  $t = (x_1 + x_2)$  or of  $x_m = m$  is assignable to more than one cell if  $t = 2$ . Cells specified by  $x_m = m$  lie on two sides of a square of  $m$  cells, there being  $(2m - 1)$  in all. If we write  $P_{12 \cdot s}$  for the probability that the sample records the unique pair of score values  $x_1$  and  $x_2$  when the number of sectors is  $s$ ,  $P_{12 \cdot m}$  for the probability that  $x$  has these two values when  $x_m = m$  and  $P_{m \cdot s}$  for the probability that  $x_m = m$  when  $s$  is the number of sectors, we thus see that

$$P_{12 \cdot s} = \frac{1}{s^2}; \quad P_{12 \cdot m} = \frac{1}{2m - 1}; \quad P_{m \cdot s} = \frac{2m - 1}{s^2}.$$

Hence in accordance with the product rule for conditional probabilities

$$P_{12 \cdot s} = P_{12 \cdot m} \cdot P_{m \cdot s}.$$

We have thus split the probability assignable to the bivariate score  $x_1, x_2$  into two factors one of which ( $P_{12 \cdot m}$ ) is independent of  $s$ ; and we might be tempted to think that we could specify



a corresponding identity  $P_{12 \cdot s} = P_{12 \cdot t} \cdot P_{t \cdot s}$  referable to the probabilities of getting the score sum  $t$  when there are  $s$  factors and getting the particular value of the bivariate score if also  $(x_1 + x_2)$  has the particular value  $t$ . Actually we cannot do so. All samples such that  $(x_1 + x_2) = t$  lie in a diagonal of  $(t - 1)$  cells if  $s \geq (t - 1)$ ; and if we knew this we might write  $P_{12 \cdot t} = (t - 1)^{-1}$  which is again independent of  $s$ . Thus there will be 4 cells in the diagonal corresponding to  $t = 5$  if  $s \geq 4$ ; but there will be only 2 cells in it if  $s = 3$ . Given  $t$  we can say that  $s \geq \frac{1}{2}t$ , e.g.  $s > 2$  if  $t = 5$ , but we cannot say that  $s \geq (t - 1)$ . The mere fact that  $t = 5$  is therefore *insufficient* to assign a unique value to the conditional probability  $P_{12 \cdot t}$ .

In the same sense we may speak of the number  $(x)$  of successes in an  $r$ -fold sample from an infinite 2-class universe as a sufficient statistic of the parameter  $p$ . We may denote by  $P_{(1 \cdot 2 \cdot 3 \dots r) \cdot p}$  the probability that the sample records successes and failures in a fixed order, there being  $r_{(x)}$  different samples so distinguishable for the particular value  $x$ . Thus we may write

$$P_{(123 \dots r) \cdot p} = p^x q^x; \quad P_{(123 \dots r) \cdot x} = \frac{1}{r_{(x)}}; \quad P_{x \cdot p} = r_{(x)} p^x q^x;$$

$$P_{(123 \dots r) \cdot p} = P_{(123 \dots r) \cdot x} \cdot P_{x \cdot p}.$$

One circumstance which gives the concept of sufficiency a peculiar importance *vis-à-vis* Fisher's approach to the problem of interval estimation is that it is not always possible to specify a sample by a statistic which is *sufficient* in his sense of the term. Since the fiducial probability distribution is in his formulation referable only to sufficient statistics and only to sufficient statistics themselves referable to continuous distributions, the fiducial theory of interval estimation is of much more limited application on its own terms than is Neyman's theory of confidence.

## 20.08 THE CLAIMS OF SMALL SAMPLE THEORY

*En passant* in 20.03 (p. 859) and more explicitly at the end of 20.04, we have had occasion to emphasise an essential difference between the Yule-Fisher interpretation of test procedure in terms of *significance* and the Neyman-Pearson-Wald approach in terms of *decision*. More explicitly, we may distinguish between a test procedure of the latter type as one designed to give a yes-or-no answer and an alternative prescription of which the only *decisive* outcome with an assignable uncertainty safeguard is the negation of a particular hypothesis. Purists of the Yule-Fisher school may therefore say that their test prescription excludes the possibility of making an *error of the second kind*. We shall now seek to clarify the limitations this renunciation imposes on the laboratory or field worker.

For three decades, we have learned to think of economy of sample size as a prior desideratum of test procedure and to envisage their applicability to small samples as the supreme merit of the type of significance test dealt within Chapter 17. Whatever else of lasting value emerges from the Neyman-Pearson concept of *test power*, it is clear that we must now re-examine any such claims without confusing what is a purely algebraic with what is wholly a logical issue. Since the practical objective of a test procedure is to arrive at a decision of some sort, we must indeed distinguish between the adequacy of a statistical technique: (a) to assign a precise uncertainty safeguard to whatever positive assertion it entitles us to make with the minimum expenditure of effort; (b) to give a decisive answer to a particular question with the utmost economy of materials. No one could now question that the class of tests prescribed by the school of R. A. Fisher are economical in terms of (a), e.g. within the same framework of initial assumptions about the structure of the parent universe the *t*-test permits us to assign a more precise uncertainty safeguard than a *c*-test to the decisive assertion that the null hypothesis is false unless the sample



is very large. That this is so, does not dispose of a possibility of more concern to the research worker. We achieve no economy by using small samples if our test procedure can assign no uncertainty safeguard to the majority of statements it leads us to make, i.e. if the overwhelming majority of permissible verdicts consistent with the choice of a false one as our null hypothesis are *unproven*.

Against the background of 20.04, we cannot disclaim the obligation to examine this possibility; and may do so without invoking any sophisticated mathematics. It will suffice if we take a back-stage view of the mechanism of the significance test for a proportionate score difference, or so-called Chi-Square test for 1 d.f., in the taxonomic domain. We shall postulate  $p_a$  and  $p_b$  as the true success rates for two treatment procedures ( $A$  and  $B$ ), denoting by  $p_{a.s}$  and  $p_{b.s}$  the corresponding success observed rates for equal ( $r$ -fold) samples. If we write for brevity  $(p_b - p_a) = d$  and  $(p_{b.s} - p_{a.s}) = d_s$ , we may then define without appreciable error for sizeable samples (e.g.  $r = 50$ ) and for values of  $p_b$  or  $p_a$  in the neighbourhood of 0.5, a square normal standard score of unit variance by the relations

$$c^2 = \frac{(d_s - d)^2}{\sigma_d^2} \quad \text{and} \quad \sigma_d^2 = \frac{p_a q_a + p_b q_b}{r} \quad (\text{i})$$

For heuristic purposes we may assume that we know the true value of  $p_a = 0.5$  for the yardstick treatment ( $A$ ), in which event

$$\sigma_d^2 = \frac{1 - 2d^2}{2r} \quad (\text{ii})$$

If we adopt the conventionally prescribed null hypothesis  $d \leq 0$ , i.e. that treatment  $B$  is no better than treatment  $A$ , we may define for  $d = 0$  our standard score in accordance with (i) and (ii) above as  $c_0 = d_s \sqrt{2r}$ . Whence for equal samples  $r = 50$ ,  $c_0 = 10d_s$ . To assign an uncertain safeguard  $\alpha \leq 0.05$  to the rejection of the hypothesis, we must make  $c = +1.64$  at the rejection level; and we shall then reject the null hypothesis only if  $d_s \geq +0.164$ .

That the probability of falsely rejecting a hypothesis which is in any event irrelevant to our practical aim does not exceed 5 per cent. will give us little reason for satisfaction if the same rejection criterion commonly leads us to an indecisive verdict when another—and maybe more relevant—hypothesis happens to be true. Taking a backstage view, we shall therefore suppose that we know treatment  $B$  to be at least 10 per cent. better than treatment  $A$ , i.e. the true hypothesis is  $d = 0.1$ . We shall thus define a standard score for  $d = 0.1$  in terms of (i) and (ii) when  $r = 50$  as

$$c_1 = \frac{10(d_s - 0.1)}{\sqrt{0.98}}.$$

We then recall that we have decided to reject the null hypothesis ( $d \leq 0$ ) if  $d_s \geq 0.164$ , i.e. if

$$c_1 \geq \frac{0.64}{\sqrt{0.98}} \quad \text{or} \quad c_1 \geq +0.65.$$

We are now in a position to answer the question we have raised above. We wish to know how often the conventional null hypothesis ( $d = 0$ ) would lead us to suspend judgment when the truth is that  $d = 0.1$ . This is simply definable in terms of the area of the normal integral of unit variance in the range from  $c_1 = -\infty$  to  $c_1 = +0.65$ , i.e.

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{0.65} e^{-\frac{1}{2}c^2} dc \simeq 0.74.$$



Thus the conventional null hypothesis test (at 5 per cent. significance level) will fail to result in a decision to reject it when the true operational advantage of treatment  $B$  is 10 per cent. in roughly 75 per cent. of the situations we shall encounter; and the choice of a more exacting criterion of rejection will merely worsen our plight. We can guarantee with higher frequency a decisive statement, i.e. rejection of the prescribed null hypothesis ( $d \leq 0$ ) for the same conditions stated ( $d = 0$  and  $p_a = 0.5$ ) with an uncertainty safeguard  $\alpha \leq 0.05$ , only if we increase  $r$  thereby diminishing our rejection criterion  $d_s$ . In short we can evade the sin of committing the error of the second kind only by incurring the risk of suspending judgment in most applications of the test procedure. If we wish to guarantee that 95 per cent. of our tests will lead us to a decisive conclusion, we have to adopt the dual test procedure of equalising the two risks ( $\alpha = 0.05 = \beta$ ) under a different name; but we can do this only by prescribing sample size in conformity with the requirements of an admissible alternative hypothesis.

The present position is therefore this. The several criteria of excellence (consistency, efficiency and sufficiency) claimed for now commonly used significance tests referable to a unique null hypothesis are economical only in a special sense with little relevance to the main concern of the research worker, i.e. to attach an uncertainty safeguard to a clear-cut decision. What is equally relevant to the contemporary revaluation of current practice is that the preference for sufficient and efficient statistics limits our choice of a null hypothesis to what a restricted range of sampling distributions can accomplish, and hence to the formulation of the null hypothesis in terms which may have no relevance to operational intention. Indeed, we may sum up the argument of 20.03–20.04 by saying that the examination of small samples in terms of the now most widely used test procedures invoking a single null hypothesis will lead to answers of only two sorts—one non-committal, the other quite often immaterial. Unless we examine the situation through the spectacles of the dual test procedure, we have no guarantee that the test performance in the overwhelming majority of situations will not lead to a non-committal answer, i.e. to no answer at all.

#### 20.09 THE SHOW MUST GO ON

In the foreword to Vol. I of *Chance and Choice* the writer expressed the view that increasing demand for instruction in statistical techniques must encourage authoritarian attitudes in higher education unless there is a more lively recognition of the need for simple exposition of the logical assumptions inherent in the mathematical derivation of statistical procedures in common use. The new concepts discussed in this chapter offer the prospect of a radical re-examination of their credentials, and a challenge to which the conscientious teacher will respond. The time for revaluation is overdue, and in one sense, therefore, it is a most encouraging sign of our time that the basic concepts of statistical theory are in the melting pot. Meanwhile, it has what may be discouraging consequences to the student confronted with contradictory assertions at the most elementary level in current text books.

Until the publication of two recent books,\* Wald's (1947) *Sequential Analysis* and Neyman's (1950) *First Course in Probability and Statistics*, much of the subject matter of this chapter was accessible only in highly abstruse mathematical publications. Inevitably, therefore, there are as yet few who have fully assimilated the logical implications of the new concepts discussed in this context. By the same token, the teacher who does not confine his or her theme to the exposition of one corpus of current dogma will find that his or her own views undergo modification under the impact of recent advances in statistical theory. This is conspicuously true of Kendall's invaluable treatise. In a rapidly and otherwise hopefully changing milieu, inconsistencies of

\* Also Feller's (1950) first volume of an *Introduction to Probability Theory and its Applications*.



statement in the writing of a book such as *Chance and Choice* have also been inescapable. For instance, the views expressed in 20.03 are inconsistent with that expressed on pp. 96-99 of Chapter 2 in Vol. I; and the idiom of pp. 212-215 of Chapter 5 in Vol. I is not consistent with the more careful statement of the same issue in 20.05.

By employing visual aids to exhibit the formal algebra of the classical theory of probability as in Vol. I and the derivation of the more widely used sampling distributions invoked by current statistical procedures as in this volume, the writer hopes that the outcome will be helpful to research workers hitherto hesitant to examine factual implications of statistical theory too long concealed by a façade of—to most of us—formidable mathematical operations; and if the method of exposition proves to be useful to teachers who are themselves seeking to formulate the credentials of statistical theory in a less authoritative temper than that of their forerunners, the effort of completing a self-imposed task will not have been wholly fruitless. None the less, he retires from the stage with the misgiving that the issues raised in the controversy between the schools of R. A. Fisher on the one hand and of Neyman, Wald and E. S. Pearson on the other will in retrospect appear to be far more challenging than the last few pages may suggest.

In retracing our steps with Neyman to the *milieu* in which the Founding Fathers set out to frame rules for the division of stakes to ensure success to the gambler who pursues any such rule consistently, we can interpret the risk of erroneous judgment only in terms consistent with a *forward look*. The risk we specify as our uncertainty safeguard is the risk of error associated only with the *entire class of statements* subsumed in a rule which we must state in advance. The reinstatement of the classical viewpoint thus deprives us of any right to associate a probability with a particular statement referable to a particular event, least of all to an event in the past. What is less obvious and has been emphasised too little in the foregoing pages is that a rule meaningfully so conceived must assign in advance the number of unit trials per game, i.e. the size of the sample. Thus the attractive algebraic properties which endow such distributions as the *t*- and the *F*- with a special claim to exactitude in one sense fail to confer on them the more exacting claims of a consistently forward look.

If we adopt a behaviourist viewpoint to the proper uses of the algebraic theory of probability, we must repudiate any concepts referable to indefinable states of mind, restricting the terms of reference of the theory in the real world of human experience to long-run frequencies of events and statements about events in situations to which the stochastic calculus has a relevance endorsed by empirical evidence. To realise all the consequences of this reorientation, we shall then need to equip ourselves with a new vocabulary; and much of the phraseology of this chapter will prove to be exceptionable if we undertake to reinterpret the legitimate scope of a calculus of judgments in an idiom consistent with the doctrine of chance in the setting of Pascal, J. Bernoulli and D'Alembert.

So far as we can now foresee, such a restatement will extensively restrict the licence to invoke stochastic considerations in situations of which we cannot with full assurance postulate randomness *sensu stricto*. It will also restrict the form our statements can take when we are dealing with choice on which we impose randomisation by recipe. If we follow to the bitter end the trail which the new American school has blazed, we may therefore have to relinquish many hitherto cherished illusions concerning what help statistical theory can offer to the research worker. Such, as yet but dimly recognised, implications of the reorientation in which they invite us to participate sufficiently accounts for the unpalatability of their views. For my part, I am content to express the hope that the reader will face the challenge; and, if convinced that a transvaluation of statistical theory is overdue, accept what limitations intellectual rectitude imposes on its claims with cheerful resignation.



## APPENDIX 1

### CHOICE OF SYMBOLS

IN this volume I have employed some symbolic conventions which the reader will not find in use elsewhere, and it is fitting to add a few comments thereon. In general, algebraic symbolism advantageously conforms to the following requirements: (i) it should be *evocative* in the sense that its form is suggestive of its meaning either by consistent acceptance of arbitrary conventions or by literal association with the meaningful content; (ii) it should be as *explicit* as necessary in the context without being more *cumbersome* than need be. Neither of these desiderata is realisable without compromise in certain situations.

For instance, some writers use  $S_n$  as shorthand for the sum of  $n$  terms and  $S_n^2$  for the sum of the squares of  $n$  terms, when the latter form at its face value would suggest the square of the sum of  $n$  terms. One might write  $S_{qn}$  or  $S_{nq}$  for the latter, employing  $q$  to indicate the quadratic term, thereby exposing oneself to misunderstanding at another level, since  $S_{nq}$  or  $S_{qn}$  might mean the sum of  $nq$  terms. The alternative  $S_{n^2}$  would be cumbersome both from a semantic and from a typographic viewpoint. In such situations, I have deliberately defined a symbol *ad hoc* (e.g.  $S_{2n}$  on p. 495), regardless of its evocative content or of established usage.

As regards (i), I have deliberately used  $n$  and  $r$  respectively for size of universe and size of sample, because the student is familiar with this convention in the domain of permutations and combinations; and I have used subscript notation more extensively than do many teachers of the elements of statistical theory because the reference is easy to recognise or to remember by recourse to the initial letter of the appropriate epithet. Thus  $M_t$  for the true value of the universe mean and  $M_s$  for a sample mean are self-explanatory, if one commonly uses the subscripts  $t$  and  $s$  in this sense. For labelling cells (p. 409) of the chessboard diagram, I have not adhered to the matrix convention  $x_{rc}$  for the score referable to the  $r$ th row and the  $c$ th column. Throughout this volume, the particular cell score  $x_{ij}$  signifies that of column  $i$  and of row  $j$ . The general term  $x_{rc}$  means a score in a column of  $r$  cells and in a row of  $c$  cells. If this offends a purist, I plead in extenuation that: (a) the chessboard is not a matrix in the ordinary sense of the term; (b) I have avoided recourse to matrix operations on the assumption that many readers with sufficient knowledge of differential calculus for understanding the book as a whole are not familiar therewith. Absolute uniformity is alas unattainable while current instruction condones  $\sin^{-1} \cdot a$  for  $\arcsin \cdot a$  in conformity with the standard convention for inverse operations, but  $\sin^n \cdot a = (\sin \cdot a)^n$ . Similarly, we still use  $\text{antilog}_b \cdot a$  where  $\log_b^{-1} \cdot a$  would be more in keeping with general usage.

With reference to (ii) above, I wish to emphasise that algebra endorses by general consent a compromise common to all communication. One cannot steer a middle course between being too vague and being too long-winded if one repudiates the right to rely on context where context suffices to clarify meaning. My extensive use of the *dot* subscript symbolism (e.g.  $E_{a \cdot bc}$ ) to distinguish operations referable to different dimensions is an attempt to avoid ambiguities which arise by leaving too much to context, and is consistent with an established convention of multiple regression and partial correlation. Some such convention is indispensable as a steering device through the maze of symbols invoked in the treatment of regression, of non-replacement sampling, and of patterns of variance analysis involving more than 2 dimensions of classification; but I have not hesitated to drop it when I regard its unwieldiness as a handicap in a sufficiently explicit context.

If I have sometimes used more explicit symbols than the context requires, it may perhaps have the advantage that it prepares the reader for their use when truly essential at a later stage.



APPENDIX 2

$$y = kx + C \quad . \quad . \quad . \quad . \quad . \quad . \quad (i)$$
$$y_3 = kx_3 + C - \epsilon_3$$

We also assume that the selected fixed values of  $x_1, x_2, x_3$  are not subject to error or, what comes to the same thing, that errors to which  $y_r$  and  $x_r$  are subject are additive. On this understanding

we wish to weight  $y_1, y_2, y_3$  so that the derived estimate ( $k_s$ ) of  $k$  is unbiased with minimum variance. Thus we initially postulate  $k_s$  as a weighted average of  $y_r$ :

$$k_s = W_1 y_1 + W_2 y_2 + W_3 y_3 \quad . \quad . \quad . \quad . \quad . \quad (\text{ii})$$

If our estimator is unbiased  $E(k_s) = k$ ,

$$\therefore W_1 E(y_1) + W_2 E(y_2) + W_3 E(y_3) = k = \sum_{r=1}^{r=3} W_r E(y_r) \quad . \quad . \quad . \quad \text{(iii)}$$

In this equation

$$W_r E(y_r) = W_r k x_r + W_r C - W_r E(\epsilon_r) \quad . \quad . \quad . \quad . \quad (\text{iv})$$

On the assumption that the true value of  $y_r$  is the mean of an indefinitely large number of observed values  $y_{r.s}$ , the mean value of the error terms in (iv) is zero, i.e.  $E(\epsilon_r) = 0$ , so that

$$\therefore \sum_{r=1}^{r=3} W_r E(y_r) = k \sum_{r=1}^{r=3} W_r x_r + C \sum_{r=1}^{r=3} W_r.$$

To satisfy (ii) and hence the condition that  $k_s$  is an unbiased estimator we must make

$$\sum_{r=1}^{r=3} W_r x_r = 1 \quad \text{and} \quad \sum_{r=1}^{r=3} W_r = 0 \quad . \quad . \quad . \quad . \quad (v)$$

More fully we may write these equations as

$$W_1x_1 + W_2x_2 + W_3x_3 = 1 \quad \text{and} \quad W_1 + W_2 + W_3 = 0 \quad . \quad . \quad (\text{vi})$$

Hence also

$$W_1x_2 + W_2x_2 + W_3x_2 = 0 = W_1x_3 + W_2x_3 + W_3x_3,$$

so that

$$W_2 = \frac{1 - W_1 x_1 + W_1 x_3}{x_2 - x_3} \quad \text{and} \quad W_3 = \frac{1 - W_1 x_1 + W_1 x_2}{x_3 - x_2}. \quad (\text{vii})$$

Equation (vii) thus defines the relations between the weights  $W_1, W_2, W_3$  in (ii) and the chosen fixed values  $x_1, x_2, x_3$  consistent with the condition that  $k_s$  is an unbiased estimator of  $k$ . To say that it has minimal variance is to say that  $V_{k.s} = E(k_s - k)^2$  is as small as possible. Now we may define the sampling variance of  $V_{k.s}$  in accordance with (ii) in terms of the variances ( $\sigma_r^2$ ) of the distribution of the values  $y_{r.s}$  of  $y$  associated with a fixed value ( $x_r$ ) of  $x$ , viz.

$$V_{k.s} = W_1^2 \sigma_1^2 + W_2^2 \sigma_2^2 + W_3^2 \sigma_3^2.$$

Our assumption is that  $y_{r \cdot s} = y_r + \epsilon_{r \cdot s}$ , i.e. the variance of  $y_{r \cdot s}$  depends only on that ( $\sigma_\epsilon^2$ ) of the error distribution which is independent of  $y_r$ . Thus  $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_\epsilon^2$ ,

$$\therefore V_{k,s} = (W_1^2 + W_2^2 + W_3^2) \sigma_\ell^2 \quad \text{. . . . . (viii)}$$

To choose our weights  $W_1, W_2, W_3$  so that  $V_{k..s}$  will be a minimum, we must satisfy the condition

$$\frac{\partial V_{k \cdot s}}{\partial W_1} = \frac{\partial V_{k \cdot s}}{\partial W_2} = \frac{\partial V_{k \cdot s}}{\partial W_3} = 0 \quad . \quad . \quad . \quad . \quad (ix)$$

We thus obtain from (viii) :

$$\frac{\partial W_1^2}{\partial W_r} + \frac{\partial W_2^2}{\partial W_r} + \frac{\partial W_3^2}{\partial W_r} = 0 \quad . \quad . \quad . \quad . \quad . \quad (x)$$

It suffices to evaluate this for  $W_r = W_1$ , so that

$$W_1 + W_2 \frac{\partial W_2}{\partial W_1} + W_3 \frac{\partial W_3}{\partial W_1} = 0 \quad . \quad . \quad . \quad . \quad . \quad (xi)$$



We derive from (vii) :

$$W_2 \frac{\partial W_2}{\partial W_1} = \frac{(1 - W_1 x_1 + W_1 x_3)(x_3 - x_1)}{(x_2 - x_3)^2};$$

$$W_3 \frac{\partial W_3}{\partial W_1} = \frac{(1 - W_1 x_1 + W_1 x_2)(x_2 - x_1)}{(x_2 - x_3)^2}.$$

On substituting these values in (xi) we get

$$(x_2 - x_3)^2 W_1 + (x_3 - x_1)(1 - W_1 x_1 + W_1 x_3) + (x_2 - x_1)(1 - W_1 x_1 + W_1 x_2) = 0,$$

$$\therefore 2W_1(x_1^2 + x_2^2 + x_3^2 - x_1 x_2 - x_1 x_3 - x_2 x_3) = 2x_1 - x_2 - x_3 \quad \text{. . . . . (xii)}$$

We can simplify (xii) by introducing  $M_x$ , the mean value of  $x_r$ , so that

$$3M_x = x_1 + x_2 + x_3 \quad \text{and} \quad 3(x_1 - M_x) = 2x_1 - x_2 - x_3;$$

$$3(x_1 - M_x)^2 + 3(x_2 - M_x)^2 + 3(x_3 - M_x)^2 = 2(x_1^2 + x_2^2 + x_3^2 - x_1 x_2 - x_1 x_3 - x_2 x_3).$$

Whence by substitution in (xii) :

$$W_1 = \frac{x_1 - M_x}{(x_1 - M_x)^2 + (x_2 - M_x)^2 + (x_3 - M_x)^2}.$$

More generally for  $n$  paired values  $(y_r, x_r)$  we may write

$$W_r = \frac{x_r - M_x}{\sum_{r=1}^{r=n} (x_r - M_x)^2}.$$

Whence by substitution in (ii) :

$$k_s = \frac{\sum_{r=1}^{r=n} y_r (x_r - M_x)}{\sum_{r=1}^{r=n} (x_r - M_x)^2} \quad \text{. . . . . (xiii)}$$

We may express this relation in another form by using the substitution

$$\sum_{r=1}^{r=n} y_r = n \cdot M_y.$$

In (xiii) above

$$y_r(x_r - M_x) = (y_r - M_y)(x_r - M_x) + M_y(x_r - M_x) \quad \text{and} \quad M_y \sum_{r=1}^{r=n} (x_r - M_x) = 0.$$

Hence we may write (xiii) in the more familiar guise :

$$k_s = \frac{\sum_{r=1}^{r=n} (y_r - M_y)(x_r - M_x)}{\sum_{r=1}^{r=n} (x_r - M_x)^2}.$$





# INDEX

(In compiling this index, key references only have been cited ; relevant cross-references are freely mentioned in the text.)

- Analysis of Covariance, 764, 771, 783
- Analysis of Variance, 407, 410, 448, 450, 455, 532
  - Additive principle in, 554
  - confidence limits, 681
  - degrees of freedom, 572, 682
  - for one criterion of classification, 569
  - for two criteria of classification, 448, 533
  - for three criteria of classification, 455, 543, 560
  - interaction, 564
  - model I and model II, 548, 731, 752
  - replication, 556, 560
  - significance tests, 669
  - variance ratio, 655, 699
- Approximations, 45
  - for factorials of large numbers, 237
  - in solution of differential equations, 48
  - in summation, 54
- Bayes' postulate, 198, 206, 853
  - theorem, 195, 849, 897
- Behrens test, 897
- Bernoulli's theorem, 133, 147
- Bernoullian universe, 510, 514, 626, 897
- Beta function, 229, 251, 256, 646, 650
- Binomial distribution, 28, 37, 110, 229, 295, 594, 605
  - histogram, 110, 115, 223
- Bivariate universe, 326, 360, 428, 475, 714
- Burette universe, 607, 620, 827
- c*-test (critical ratio), 127, 128, 148, 187, 192, 203, 213, 303, 313, 323, 691, 699, 906
- Central difference, 116, 120, 230
- Chi-Square distribution, 217, 257, 263, 427, 612, 644, 663, 665, 667, 670, 828, 833
- Classification, 60, 68
  - criteria of, in analysis of variance, 533
  - manifold, 428, 804, 828
- Co-moments, 808
- Co-prime samples, 169, 340
- Concomitant variation (*see* correlation), 326, 360, 475, 482, 527, 528
- Concurrence, 326, 369, 384, 389, 400, 747, 750, 784
- Confidence limits, 211, 213, 219, 695, 887
  - theory, 859, 885, 897
- Consequence, 326, 369, 384, 389, 400, 748, 750, 784
- Contingency grid, 429
- Correlation, 326
  - in factor analysis, 784
  - grid, 358, 429
- Correlation—*cont.*
  - and linear regression, 383, 388, 528
  - multiple, 752
  - partial, 390, 483, 527
  - product-moment, 344, 349, 353
  - rank, 331, 350
  - ratio, 379, 441, 706
  - tautologies of, 441
  - umpire-bonus model in, 360
  - universe, 476, 477
- Covariance, 344, 360
  - addition of, 459
  - analysis of, 764, 771, 783
- Curve-fitting, 110, 580, 595
  - by method of moments, 232, 254
  - and regression, 712, 746
- Decision test, 848
- Design of Experiments, 558, 783
- Difference distributions, 143, 153, 179, 183, 194, 271, 290, 611, 825
  - equation, 119
- Diophantine equation, 171
- Discriminant function, 759
- Distribution function, 230
- Double dichotomy, 880
- E*-notation, 434, 453, 459
- Efficiency, 217, 294, 746
- Electivity, 75, 81, 91, 101
- Errors of first and second kind, 862
- Expectation fit, 581
- F*-distribution, 653, 655, 699, 710, 740, 773, 783
- Factor Analysis, 784, 802
  - attenuation, 791
  - factor pattern, 785, 793, 795, 798
  - hierarchical principle, 785
  - reliability, 791
  - saturation, 792
  - umpire-bonus model in, 484, 487
- Fiducial distributions, 897
  - limits, 213, 219
- Figurate series, 13
  - in sampling, 65
  - in summation, 462
- Fixed *A*-set, 714, 750
- Frequency grid, 407, 429, 432
  - histogram, 110, 581, 636
  - proportionate and relative, 101, 543



- Gamma function, 229, 246, 251, 256, 589, 646, 660  
 tabulation, 659
- Gregory's formula, 33, 35, 52, 235
- Grid,  
   contingency, 429  
   correlation, 358, 429  
   frequency, 407, 429  
   independence, 439, 465  
   regression in, 442  
   score, 407, 410, 429, 448  
   tautologies of, 428  
   types of, 428
- Half-interval correction, 115, 116, 127
- Homogeneity, 453, 541, 897  
   criteria of, 413, 543
- Homoscedasticity, 384, 480, 714, 721
- Hypergeometric distribution, 139, 230, 804
- Independence, 326, 331, 349, 635  
   condition, 665  
   grid, 439, 465
- Integration, 50, 234
- Interval Estimation, 886, 897
- Inverse probability, 199
- Leibnitz' rule, 587
- Lexis models, 394, 508
- Likelihood, 197
- Maclaurin's theorem, 35, 46, 234, 265
- Mean deviation, 231, 250
- Method of least squares, 713, 746, Appendix II
- Modular likelihood, 204, 340, 845, 863
- Moments, 229, 579  
   as descriptive parameters, 231  
   as gamma functions, 250  
   derivation of, 582  
   factorial, 591, 804  
   generating functions, 264, 465, 582, 601  
   of Bernoullian universe, 626  
   of binomial distribution, 594  
   of chi-square distribution, 833  
   of difference distribution, 271, 611  
   of distribution of the mean, 602  
   of normal distribution, 595  
   of Poisson distribution, 592  
   of rectangular distribution, 593  
   of score-sum distribution, 807
- Multinomial theorem, 39, 42, 79, 809
- Necessary and Sufficient Conditions, 528
- Non-replacement distribution, 137, 490, 814, 825
- Normal distribution, 115, 127, 139, 230, 250, 279, 294, 317, 427, 595  
   approximations to, 616
- Null hypothesis, 96, 105, 143, 153, 187, 205, 207, 840, 859
- Ordinate fit, 581
- Orthogonal transformation, 520, 674
- Paired differences, 312, 630, 691  
   *c*-test for, 313  
   *t*-test for, 313, 704
- Pascal's triangle, 25, 66
- Pearson system, 254, 427, 634, 646
- Pearson's coefficients, 232, 254, 342, 593, 604, 634, 821  
   of binomial distribution, 605  
   of burette universe, 607  
   of Poisson distribution, 605  
   of rectangular distribution, 605
- Poisson distribution, 136, 223, 403  
   moments, 592  
   Pearson coefficients, 605
- Posterior probability, 197, 852
- Prior probability, 195, 197, 206, 852, 867
- Probability density, 183, 636  
   generating function, 465, 467  
   integral, 127
- Probable error, 204, 659
- Quality Control, 842
- Rectangular distribution, 255, 261, 280  
   moments, 593  
   Pearson's coefficients, 605
- Regression, 377, 528, 712, 764  
   as standardising device, 767  
   coefficients, 383, 388, 442  
   equation, computation for, 726  
   estimates, 728, 740, 756  
   linear, 377, 289, 442, 485, 492, 501, 505, 527, 765  
   multiple, 752
- Rigour, 580
- Sampling, 58, 394, 428  
   classified, 68  
   distributions, 148, 153, 166, 535, 634, 642  
   from different universes, 294, 625  
   models, 59  
   randomisation, 62, 94  
   restrictive and repetitive, 71  
   size, and significance, 218, 906  
   size, as a source of variation, 517  
   without replacement, 71, 508, 804, 814
- Score-grid, 407, 410, 430  
   summarising indices, 448  
   symbolism, 448  
   tautologies, 448, 532  
   three-dimensional, 452
- Scoring, taxonomic and representative, 194, 275, 490, 494, 533, 536, 814, 828
- Sequential ratio, 849, 854, 880



- Significance, 104, 105, 108, 195, 222, 840  
    and sample size, 218, 906  
    test, 105, 148, 187, 203, 275, 579, 634, 848, 906  
        for analysis of variance, 669, 706  
        for regression estimates, 734, 740, 759  
Small sample theory, 906  
Standard error, 304  
Standard score (*see* Critical ratio), 128, 526, 598, 830  
Standardisation, 767  
Statistical estimation, 307, 840, 885, 897  
    inference, 840  
        conditional and unconditional, 841, 885  
    inspection, 842  
Stirling's formula, 46, 47, 234, 240, 242  
Stochastic credibility, 872  
Stringency, 872  
Sufficiency, 528, 904  
Summarisation,  
    algorithms of, 427  
    correlation coefficient in, 351, 360, 530  
Symmetrical distributions, 472  
  
*t*-test, 325, 427, 655, 665, 692, 783, 906  
    for paired differences, 704  
Tchebychev's theorem, 134  
  
Test power, 873  
    procedures, 840, 859, 871, 880, 906  
Therapeutic trial, 312, 840, 859, 874, 877  
  
Umpire-bonus model, 360, 481, 520  
    algebraic properties, 385  
    and factor analysis, 786  
    and partial correlation, 390  
    unrestricted, 372  
Uncertainty safeguard, 840, 852  
Unit sampling distribution, 278, 536  
    univariate and bivariate, 477  
  
Vandermonde's theorem, 37, 299  
Vanishing triangle, 31  
Variance, 120, 132, 230  
    analysis of, 532, 669  
    balance sheet, 411, 535, 547, 560  
    explained and unexplained, 369, 370, 383, 388, 389, 400, 500, 514  
    interclass and intraclass, 369, 395  
    of a difference, 174  
    of non-replacement distribution, 137  
    of regression coefficient, 734  
    partition of, 394  
Vector likelihood, 204, 340, 845, 863  
  
Wallis product, 240